ЭЛИЕЗЕР ЮДКОВСКИЙ
НЕЙТ СОАРЕС

ЕСЛИ КТО-ТО ЕГО СОЗДАСТ — ВСЕ ПОГИБНУТ

ELIEZER YUDKOWSKY AND NATE SOARES

IF ANYONE BUILDS IT, EVERYONE DIES

WHY SUPERHUMAN AI WOULD KILL US ALL

ЭЛИЕЗЕР ЮДКОВСКИЙ, НЕЙТ СОАРЕС

ЕСЛИ КТО-ТО ЕГО СОЗДАСТ — ВСЕ ПОГИБНУТ

ПОЧЕМУ СВЕРХЧЕЛОВЕЧЕСКИЙ **ИИ** УНИЧТОЖИТ НАС ВСЕХ

Перевод с английского
Евгения Поникарова

издательство **АСТ**
Москва

УДК 004.8
ББК 32.813
Ю16

Художественное оформление и макет Андрея Бондаренко

Юдковский, Элиезер.
Ю16 Если кто-то его создаст — все погибнут. Почему сверхчеловеческий ИИ уничтожит нас всех / Элиезер Юдковский, Нейт Соарес; пер. с англ. Е. Поникарова. — Москва : Издательство АСТ : CORPUS, 2026. — 256 с.

ISBN 978-5-17-182679-6

В 2023 году сотни видных деятелей в области искусственного интеллекта подписали открытое письмо, предупреждающее о серьезной угрозе для человечества. Но компании и страны все равно спешат создать машины умнее человека. И мир совершенно не готов к тому, что произойдет дальше.
Несколько десятилетий двое специалистов, подписавших то письмо, — Элиезер Юдковский и Нейт Соарес — изучали, как сверхинтеллект будет мыслить и вести себя. Их исследования показывают: конфликт продвинутых ИИ-моделей и человечества неизбежен. И нам не победить. Как и зачем искусственный сверхинтеллект уничтожит наш вид? Что необходимо сделать человечеству для выживания? Авторы подробно, доступно и образно обосновывают свою точку зрения и призывают каждого человека на планете задуматься об экзистенциальной угрозе, нависшей над всеми нами. Пока еще не поздно.

УДК 004.8
ББК 32.813

ISBN 978-5-17-182679-6

© 2025 by Eliezer Yudkowsky and Nate Soares
© Е. Поникаров, перевод на русский язык, 2026
© А. Бондаренко, художественное оформление, макет, 2026
© ООО "Издательство АСТ", 2026
Издательство CORPUS ®

Содержание

Введение. Сложные прогнозы и простые прогнозы. 9

ЧАСТЬ I
Нечеловеческий разум

ГЛАВА 1 Особая сила человечества. 23
ГЛАВА 2 Выращено, а не сконструировано 35
ГЛАВА 3 Обучая хотеть. 50
ГЛАВА 4 Вы получаете не то, чему обучаете 60
ГЛАВА 5 Его любимые вещи 82
ГЛАВА 6 Мы проиграем. 98

ЧАСТЬ II
Один сценарий вымирания

ГЛАВА 7 Пробуждение. 121
ГЛАВА 8 Экспансия 135
ГЛАВА 9 Вознесение 155
Кода 161

ЧАСТЬ III
Лицом к лицу с проблемой

ГЛАВА 10 Проклятая проблема 165

ГЛАВА 11 Алхимия, а не наука 182

ГЛАВА 12 Не хочу показаться паникером 198

ГЛАВА 13 Остановите всё 212

ГЛАВА 14 Пока есть жизнь, есть и надежда 223

Заключительные слова 235

Благодарности 237

Комментарии 239

Всем людям, умершим в ходе
долгой истории нашего вида,
всем, кто еще жив,
и всем детям, которые когда-нибудь
могут появиться

ВВЕДЕНИЕ

Сложные прогнозы и простые прогнозы

“Снижение риска вымирания человечества из-за искусственного интеллекта должно стать глобальным приоритетом — наравне с пандемиями, ядерной войной и другими угрозами глобального масштаба”.

В начале 2023 года сотни ученых, работающих в сфере искусственного интеллекта, подписали открытое письмо, содержавшее одну эту фразу. Среди тех, кто поставил подпись, были весьма титулованные специалисты в этой области, в частности нобелевский лауреат Джеффри Хинтон и Йошуа Бенджио, получившие премию Тьюринга за создание глубокого обучения.

Мы, Элиезер Юдковский и Нейт Соарес, также подписали это письмо, хотя и сочли его формулировку слишком сдержанной.

Ни нас, ни других подписантов не беспокоят ИИ-модели образца 2023 года. Нас не беспокоят и те модели, что существуют на момент написания этих строк — в начале 2025 года. Современные ИИ-модели все еще кажутся поверхностными в каком-то глубинном смысле, который сложно описать. У них есть ограничения: например, неспособность формировать новые долговременные воспоминания. Эти недостатки пока что не позволяют им проводить серьезные на-

учные исследования или заменять людей на рабочих местах в сколько-нибудь значимом количестве*.

Наше беспокойство вызывает то, что появится дальше: машинный интеллект, который будет по-настоящему разумным — разумнее любого человека и даже всего человечества в целом. Нас тревожит ИИ, который превзойдет человеческую способность мыслить, обобщать опыт, решать научные задачи, изобретать новые технологии, планировать, вырабатывать стратегии, просчитывать ходы и совершенствовать себя. Когда он превзойдет человека почти во всех интеллектуальных задачах, такой ИИ можно будет назвать искусственным сверхинтеллектом.

Пока еще ИИ-модели не достигли этого уровня. Но сегодня они умнее, чем в 2023 году, и значительно умнее, чем в 2019-м. Исследования в области искусственного интеллекта приводили к одному скачку за другим: в 2012, 2016, 2020, 2022 и 2024 годах**. Мы не знаем, выдохнется ли этот прогресс, взяв паузу на какое-то время, пока не появятся новые методы и технологии. Мы не знаем, сколько таких скачков отделяет нас от момента, когда ИИ станет угрозой уровня вымирания, о которой предупреждали авторы открытого письма. Однако история раз за разом показывает: исследователи в области ИИ изобретают новые методы и преодолевают старые препятствия. Зачастую прогресс идет удивительно быстрыми темпами. В 2015 году большинство специалистов по компьютерным наукам сказали бы вам, что собеседник уровня *ChatGPT* появится только лет через тридцать или даже пятьдесят.

* Авторы благоразумно указали, что пишут эти строки в начале 2025 года. Сегодня, буквально через год, оба процесса уже вовсю начинают происходить, хотя проблема долгосрочной памяти у языковых моделей еще не вполне решена. *(Прим. науч. ред.)*

** В 2012 году нейронная сеть *AlexNet* начала справляться с проблемой распознавания изображений. В 2016-м программа *AlphaGo* победила лучшего игрока в го. В 2020-м появилась (чисто предиктивная) языковая модель GPT-3. В 2022-м появился (широко используемый) чат-бот *ChatGPT*. В 2024-м рассуждающие модели начали хорошо писать код и гораздо лучше решать математические задачи и визуальные головоломки.

Мы не знаем, когда появится искусственный сверхинтеллект, но согласны, что эта проблема должна иметь глобальный приоритет. Более того, мы считаем, что вышеуказанное открытое письмо сильно недооценивает серьезность проблемы.

Нас пригласили подписать то письмо из одной фразы, поскольку мы руководим Институтом исследований машинного интеллекта (MIRI). Эта некоммерческая организация работает над вопросами, связанными с машинным сверхинтеллектом, с 2001 года, то есть она начала задаваться этими вопросами задолго до того, как они стали широко обсуждаться и привлекать финансирование. Если говорить упрощенно: среди тех немногих, кто следит за этой темой десятилетиями, принято считать, что именно MIRI работает над ней дольше всех. Один из нас, Юдковский, — основатель MIRI; другой, Соарес, — его нынешний президент.

MIRI — первая организованная группа, которая заявила: "В какой-то момент появится искусственный сверхинтеллект, и это событие представляется чрезвычайно важным. Возможно, будет технически сложно направить работу сверхинтеллекта в такое русло, чтобы он помогал человечеству, а не вредил ему. Не лучше ли приступить к решению этой задачи прямо сейчас, а не ждать, пока ситуация превратится в масштабный кризис?"

Но начинали мы не с этого. В 2000 году Юдковский попытался создать машинный сверхинтеллект. В 2001-м он осознал, что тот не обязательно окажется дружественным. А в 2003-м он понял, что эта проблема сложна.

Первые два десятилетия своего существования MIRI был институтом технических исследований, практически не участвовавшим в политике. Организация в основном проводила семинары для заинтересованных ученых и давала пристанище нескольким многообещающим исследователям. Мы пытались разработать математический аппарат для понимания и формирования сверхчеловеческого машинного интеллекта, а также для предсказания, что может пойти не так.

Деятельность MIRI имела и непрямые последствия, к которым мы теперь относимся неоднозначно или с сожалением. На одной из организованных нами конференций мы познакомили Демиса Хассабиса и Шейна Легга, основателей компании, которая сейчас называется *Google DeepMind*, с первым их крупным инвестором. А Сэм Альтман, генеральный директор *OpenAI*, однажды заявил, что Юдковский "привил многим из нас интерес к ОИИ*" и "сыграл ключевую роль в решении запустить *OpenAI*"**.

История MIRI сложна, но наши взаимоотношения со всей этой сферой можно описать так: за несколько лет до появления нынешних ИИ-компаний любой, кто хотел, невзирая на риск вымирания, создавать по-настоящему умную модель ИИ, должен был сперва отмахнуться от наших предупреждений.

Позже, когда сфера ИИ начала набирать обороты, мы с тревогой наблюдали, как некоторые люди, основывающие новые компании, заговорили об искусственном сверхинтеллекте как об источнике огромной чудесной силы. Силы, которую они, как им казалось, сумеют контролировать. По мнению многих из них, главная опасность заключалась в том, что искусственный сверхинтеллект может "оказаться" не у тех людей. Они рассуждали о необходимости выиграть "гонку вооружений в области ИИ". А вот о том, что искусственный сверхинтеллект "окажется" у самого же искусственного сверхинтеллекта — то есть единственным победителем в гонке вооружений в области ИИ будет сам искусственный сверхинтеллект, — основатели компаний не говорили.

Мы видели, что возможности ИИ росли крайне быстро.

* Термин ОИИ (общий искусственный интеллект) предназначен для того, чтобы отличать ИИ, который интуитивно кажется "действительно умным", от узкоспециализированных ИИ-моделей прошлого. В нашей книге мы избегаем этого термина, поскольку специалисты расходятся во мнениях, что же именно он подразумевает в свете появления ИИ-моделей, подобных *ChatGPT*.

** Если это правда, то произошло это вопреки мнению Юдковского: он считал, что *OpenAI* — ужасная идея.

Мы видели, что та область исследований, где работали мы сами, — ставящая своей целью понять ИИ и, возможно, не дать ситуации пойти по плохому сценарию — развивалась намного, *намного* медленнее.

Безоглядное рвение ИИ-компаний к созданию сверхчеловеческого искусственного интеллекта — их стремление создать его как можно быстрее, опередив конкурентов, — стало выглядеть для нас как гонка по нисходящей. Индустрия неслась под откос: ситуация могла бы войти в учебники в качестве примера, как *не* надо вести разработки (правда, написать такой анализ было бы уже некому).

Мы уже не верили, что человечество сумеет найти выход из катастрофической ситуации с помощью инженерных решений и исследований. Не в таких условиях. Не к нужному сроку.

Мы сочли наши предыдущие усилия провальными, свернули бóльшую часть исследований MIRI и направили ресурсы института на донесение одной-единственной мысли — предупреждения, лежащего в основе этой книги:

> Если какая-либо компания или группа
> на планете создаст искусственный
> сверхинтеллект, используя что-либо,
> даже отдаленно похожее
> на современные технологии,
> и основываясь на понимании ИИ,
> даже отдаленно похожем на нынешнее,
> то погибнут все люди на всей Земле.

Это не гипербола. Мы не преувеличиваем ради эффекта. Мы считаем, что это самая прямая экстраполяция современных знаний, данных и институционального поведения в области искусственного интеллекта.

В этой книге мы излагаем наши доводы в надежде сплотить достаточное число как лиц, принимающих важные ре-

шения, так и обычных людей — чтобы они отнеслись к ИИ серьезно. Исход по умолчанию летален, однако ситуация пока еще не безнадежна: машинного сверхинтеллекта пока не существует, и его создание все еще можно предотвратить.

Как можно быть в чем-то уверенным, когда дело касается искусственного интеллекта? Известный афоризм гласит: "Предсказывать очень трудно, особенно будущее". Бóльшая часть того, что мы хотели бы знать о будущем, на самом деле не поддается прогнозированию. Например, мы не можем назвать вам выигрышные номера лотереи на следующей неделе. Все наборы чисел выглядят одинаково вероятными.

Но некоторые факты о будущем действительно предсказуемы. Предположим, завтра вы купите лотерейный билет. Мы не знаем, какими изощренными теориями или уловками вы воспользуетесь при выборе чисел, и не знаем, какие номера выпадут. Но вся эта неопределенность сводится к весьма надежному прогнозу: скорее всего, вы не выиграете. Аналогично, если бросить кубик льда в стакан с горячей водой, невозможно предсказать, где окажется каждая конкретная молекула десять минут спустя. Однако вся эта неопределенность складывается в прогноз, близкий к достоверному: кубик льда растает. По такому принципу работает существенная часть физики: мы знаем, куда ведут почти все пути, даже если не можем просчитать, какой именно путь будет выбран.

Одни аспекты будущего возможно предсказать, имея нужные знания и прикладывая усилия; другие же предсказать почти невозможно. Компетентная футурология строится на понимании этой разницы.

История учит нас, что относительно легко прогнозировать будущее следующим образом: вы осознаете, что нечто выглядит теоретически возможным в соответствии с законами физики, и предсказываете, что со временем кто-нибудь это сделает. Полеты аппаратов тяжелее воздуха, оружие,

высвобождающее ядерную энергию, ракеты, летящие на Луну с человеком на борту, — все это было предсказано заранее, и на то имелись разумные причины, несмотря на возражения скептиков, которые с умным видом изрекали, что раз этого еще не произошло, то, вероятно, не произойдет никогда. Люди, привязывавшие крылья к рукам и прыгавшие с холмов, выглядели крайне глупо, они калечились и терпели неудачи, их высмеивали современники — однако это не остановило братьев Райт, которые поняли, как нужно летать.

Но куда более сложная задача — предсказать, *когда* именно появится та или иная технология. Люди могут утверждать, что до появления технологии осталось года два, тогда как на самом деле — все пятьдесят, или заявляют про пятьдесят лет, когда на самом деле — всего два года, причем они сами же эту технологию и создадут. "Человек не будет летать еще тысячу лет", — бросил Уилбур Райт Орвиллу в 1901 году, когда братьев измучили неудачи при испытаниях безмоторного планера. Два года спустя, в 1903-м, аэроплан братьев Райт поднялся в воздух.

Успешное прогнозирование заключается не в том, чтобы проявить достаточно ума и предсказать те детали, которые обычно предсказанию не поддаются. И не в том, чтобы придумать полную историю того, что произойдет в будущем, а затем чудесным образом оказаться правым. Дело в поиске тех аспектов будущего, прогнозировать которые просто, если смотреть под правильным углом.

Мы не знаем, когда наступит конец света, если люди и страны ничего не изменят в своем отношении к искусственному интеллекту. Мы не знаем, какие заголовки об ИИ появятся через два года или десять лет, и даже не знаем, есть ли у нас в запасе эти десять лет. Мы не претендуем на звание настолько умных, что якобы можем предсказывать по-настоящему труднопредсказуемые вещи. Нам просто кажется, что один конкретный аспект будущего — что произойдет со всеми людьми и всем, что им дорого, если сверхинтеллект будет создан в ближайшее время, — спрогнозировать вполне

легко, если воспользоваться имеющимися знаниями и аккуратными рассуждениями.

На первый взгляд гибель человечества из-за сверхчеловеческого ИИ не выглядит логичным прогнозом. Но как раз для этого и написана наша книга. Точно так же как арифметика нужна, чтобы рассчитать шансы на победу в лотерее, а термодинамика — чтобы объяснить, почему ледяной кубик предсказуемо растает, необходима определенная подготовка, чтобы понять, почему искусственный интеллект несет угрозу неизбежного вымирания человечества. И стоит заложить такой фундамент, как прогноз исхода нынешней траектории становится удручающе, до ужаса очевидным.

Даже перед лицом сверхчеловеческого машинного интеллекта может возникнуть соблазн поверить, будто мир останется таким же, каким он был в последние десятилетия нашей относительно короткой жизни. Нам трудно осознать, что всего несколько столетий назад цивилизация кардинально отличалась от нынешней. Несколько тысячелетий назад цивилизация только зарождалась. Несколько миллионов лет назад не существовало людей. Миллиард лет назад только появлялись специализированные клетки.

Историческая перспектива помогает нам осознать то, что трудно увидеть из-за короткой продолжительности нашей жизни: природа допускает потрясения. Природа допускает бедствия. Природа допускает, что мир никогда не будет прежним.

Некогда — 2,5 миллиарда лет назад — произошло событие, которое биологи называют кислородной катастрофой: новая форма жизни научилась использовать энергию солнечного света для извлечения ценного углерода из воздуха. В качестве отходов жизнедеятельности эти организмы выделяли химически активное вещество, ядовитое для большинства существовавших тогда форм жизни: сейчас мы называем его

“кислород”. Это вещество постепенно накапливалось в атмосфере. Основная часть живых существ, включая и большинство бактерий, выделявших этот самый кислород, не сумели справиться с его химической активностью и погибли. Лишь немногие удачливые линии клеток приспособились и со временем эволюционировали в организмы, использующие кислород как топливо. Но к прежней норме возврата не было. Мир уже никогда не был прежним.

Некогда континенты были бесплодными камнями. Затем, по эволюционным меркам — в мгновение ока, они покрылись ковром растительности. И вскоре в лесах забурлила жизнь. Мир уже не был прежним.

Некогда люди одомашнили пшеницу и ячмень. Спустя ничтожный с эволюционной точки зрения миг они начали строить цивилизации. Мир уже не был прежним.

Некогда — в 1930-е годы — появились тревожные признаки того, что над некоторыми семьями в Германии нависла опасность. Немногие уехали заблаговременно, большинство осталось. Затем нацистское правительство лишило их гражданства и паспортов, сильно затруднив отъезд. Через несколько лет немецких евреев, цыган и других людей начали хватать и отправлять в лагеря смерти. Воспоминания выживших свидетельствуют, что многие семьи остались не потому, что не видели тревожных знаков, а потому, что верили: жизнь вернется к норме прежде, чем все зайдет слишком далеко[1].

Некогда человечество стояло на пороге создания искусственного сверхинтеллекта…

Нормальность всегда заканчивается. Это не значит, что ей на смену неизбежно придет что-то худшее. Иногда — да, иногда — нет, и порой это зависит от наших действий. Однако надежды, что ничто слишком плохое просто не может случиться, как правило, не помогают.

Люди обладают способностью управлять будущим с помощью своего интеллекта. Но эта способность работает, *только если мы ее используем* — если делаем то, что дол-

жны, тогда, когда это нужно. В противном случае интеллект не имеет силы. Он или работает, изменяя наши действия, или не работает вовсе.

Грядущие годы станут для всего человечества испытанием не на жизнь, а на смерть. Своей книгой мы надеемся вдохновить отдельных людей и целые страны достойно ответить на этот вызов.

В следующих главах мы изложим научные основания нашего беспокойства, обсудим порочные стимулы, движущие современной индустрией ИИ, и объясним, почему ситуация еще плачевнее, чем кажется. Мы в доступной форме покажем недостатки современного машинного обучения и объясним, почему нынешние методы ведут к созданию искусственного интеллекта, уничтожающего мир, а не улучшающего его.

В части I мы излагаем суть проблемы, отвечая на ряд вопросов. Что такое интеллект? Как создают современные ИИ-модели и почему их так трудно понять? Могут ли они иметь желания? Появятся ли эти желания? Если да, то *чего* ИИ-модели захотят и почему они захотят нас убить? Как они нас убьют? В итоге мы прогнозируем появление ИИ-моделей, не испытывающих к нам ненависти, но обладающих странными и чуждыми предпочтениями, следовать которым они будут до вымирания человечества.

В части II мы собираем воедино все эти аспекты и рассказываем историю о модели ИИ, которая уничтожает мир, очень похожий на наш. Эта история — не предсказание, поскольку точный путь, по которому пойдет будущее, спрогнозировать сложно. Единственная часть этой истории, которая служит предсказанием, — ее финал, и это предсказание сбудется, если мы позволим подобной истории начаться.

В части III мы оцениваем сложность вызова, стоящего перед человечеством, и рассматриваем принимаемые сейчас меры. Насколько хорошо компании, занимающиеся ИИ,

справляются с этой проблемой? Почему мир уделяет проблеме так мало внимания? Что наше общество может сделать по-другому, если достаточное количество людей решит *не* умирать? Что потребуется, чтобы Земля *не* создала машинный сверхинтеллект?

На сайте *IfAnyoneBuildsIt.com* подготовлено онлайн-дополнение к этой книге. В конце каждой главы вы найдете URL-адрес дополнительных материалов к ней. Это выглядит так:

IfAnyoneBuildsIt.com/intro

У людей есть самые разные, зачастую противоположные представления об искусственном интеллекте, и за эти годы мы услышали множество вопросов и возражений, проистекающих из широкого спектра предположений и точек зрения. В дополнительные материалы мы включили оговорки, тонкости и часто задаваемые вопросы, а также основные теоретические положения и развернутые аргументы, которые сделали бы эту книгу в несколько раз толще и гораздо менее доступной для понимания. Если после прочтения какой-либо главы у вас возникнут возражения, мы призываем вас перейти к онлайн-материалам.

Многие главы начинаются с притч — историй, которые, мы надеемся, помогут донести определенные мысли в более простой форме. А еще они привносят немного несерьезности в эту тяжелую тему. Тут мы следуем очень старой традиции — возможно, более древней, чем человеческий вид в его нынешней форме, — традиции смеяться в лицо смерти.

Признаться, в этой книге мало хороших новостей. Но мы и не утверждаем, что вы обречены. Искусственного сверхинтеллекта пока не существует. Человечество все еще может принять решение не создавать его.

В 1950-х годах многие ожидали, что между ведущими мировыми державами начнется ядерная война. С учетом предыдущей истории человеческих конфликтов причин для пессимизма хватало. Тем не менее ядерной войны до сих пор не случилось. И не потому, что ядерные бомбы оказались чистой научной фантастикой и не воплотились в реальность, а потому, что люди целенаправленно выстраивали устойчивые системы, призванные не допускать ядерных войн. Человечество пошло на это, поскольку мировые лидеры понимали: в случае ядерной войны плохо придется и им самим, и народам их стран.

Плохо придется им и в том случае, если кто-либо где-либо на планете создаст машинный сверхинтеллект. *Никому* не хочется умирать вместе со своей семьей и друзьями, со своей страной и ее детьми.

Остановить продолжающуюся эскалацию ИИ-технологий, взять под контроль аппаратное обеспечение, используемое для создания все более мощных моделей искусственного интеллекта, — непростая задача для современного мира. Но чтобы остановить дальнейшее наращивание возможностей ИИ, потребуется гораздо меньше усилий, чем, скажем, для победы во Второй мировой войне. Для пробуждения воли к жизни достаточно, чтобы некоторые страны, их руководители и избиратели, осознали, что люди стоят у края пропасти — возможно, на ничтожном расстоянии, пусть даже его трудно оценить.

Задача не из легких, но мы еще не погибли. Достоинство человека и достоинство всего человечества обязывают нас бороться.

Пока есть жизнь, есть и надежда.

ЧАСТЬ I
Нечеловеческий разум

ГЛАВА 1

Особая сила человечества

Представьте себе, — хотя, конечно, ничего подобного никогда не происходило, ибо это всего лишь притча, — что биологическая жизнь на Земле стала результатом игры богов. Что был бог-тигр, создавший тигров, бог-секвойя, создавший секвойи, и боги разных видов рыб и бактерий. Вообразите, что боги-игроки соперничали друг с другом, добиваясь доминирования своих видов, пока эти формы жизни пресмыкались внизу.

Представьте, что примерно два миллиона лет назад некий непримечательный бог-гоминид окинул взглядом огромное — во всю планету — игровое поле.

— Мне понадобится еще несколько ходов, — сказал он, — но, думаю, результат игры предрешен.

Воцарилось недоуменное молчание: боги в замешательстве разглядывали игровое поле, пытаясь понять, что они упустили.

Бог-скорпион спросил:

— Как это? У твоих гоминидов нет ни брони, ни когтей, ни яда.

— У них есть мозг, — ответил бог-гоминид.

— Я нашлю на них болезнь, и они умрут, — фыркнул бог-оспа[1].

— Это ненадолго, — парировал бог-гоминид. — Тебе быстро придет конец, Оспа, — как только их мозг научится с тобой бороться.

— У них даже не самый большой мозг! — воскликнул бог-кит.

— Дело не только в размере, — пояснил бог-гоминид. — Важна еще и *конструкция* их мозга. Дайте им два миллиона лет, и они ступят на луну своей планеты.

— Что-то я не вижу, где в метаболизме этого существа производится ракетное топливо, — сказал бог-секвойя. — Нельзя силой мысли выйти на орбиту. В какой-то момент твоему виду нужно развить метаболизм, который будет очищать ракетное топливо. Самим организмам придется стать довольно крупными, в идеале — высокими и узкими, с твердой внешней оболочкой, чтобы не раздуться и не умереть в космическом вакууме. Как бы усердно ни умел думать твой вид, он останется прикованным к земле, просто усердно думающим.

— Некоторые из нас играют в эту игру уже миллиарды лет, — вступил в разговор бог-бактерия, покосившись на бога-гоминида. — И до сих пор мозг не давал такие уж большие преимущества.

— И тем не менее, — ответствовал бог-гоминид.

А вот как на самом деле развивалась история нашего вида. Люди обзавелись мозгом — необычайно большим для животного такого размера. Они укротили огонь, занялись сельским хозяйством и научились плавить железо. Через поразительно короткое по меркам биологической эволюции время люди высадились на Луну, хотя наш метаболизм не умеет очищать ракетное топливо, а наша кожа не приспособлена к космическому вакууму.

Другие виды на Земле рождаются со специализированными навыками: пчелы строят ульи, бобры возводят плотины. Человек смотрит на плотину бобра и выясняет, как она устроена. Мы освоили строительство плотин, хотя эти знания не были заложены в наших генах. Теперь мы перекрываем реки плотинами, как бобры, строим себе дома, как пчелы, плетем сети из нитей, как пауки. Кроме того, мы со-

здаем электростанции и космические ракеты, чего не делает ни один другой вид.

Люди умеют делать то, чего никогда не делали их предки и чего не могут делать другие животные, благодаря качеству, которое иногда называют "интеллектом". Бóльшая часть наших способностей не запрограммирована в наших генах — просто мы наблюдали, пробовали, запоминали, обобщали и в итоге добивались результата.

Способность к обучению присуща не только людям. Мозг мыши может научиться ориентироваться в лабиринте. Но мы способны делать больше и на более высоком уровне. Можем изучать химические процессы и находить более дешевые удобрения. Можем ставить сложные эксперименты, постигать законы физики и изобретать спутники. Можем даже сажать мышей в лабиринты и изучать, как *они* обучаются.

Человеческий мозг способен прокладывать более разнообразные пути через более широкий срез реальности, нежели мозг любого другого животного. В этом и заключается наша особая сила.

Как работает эта особая сила? Что именно она делает и каким образом?

На наш взгляд*, интеллект сводится к двум фундаментальным типам работы: по *прогнозированию* мира и по *управлению* им.

"Прогнозирование" — это угадывание того, что вы увидите (или что услышите, или к чему прикоснетесь), еще до самого момента восприятия. По дороге в аэропорт ваш мозг успешно справляется с задачей прогнозирования всякий раз, когда вы предугадываете, что сигнал светофора станет желтым или что водитель впереди нажмет на тормоз.

* Эта точка зрения подкрепляется определенной теорией, которую мы обсуждаем в онлайн-материалах. Но вообще мы не собираемся слишком зацикливаться на определениях. Если удар молнии поджег лес вокруг вас, то вам не спастись, просто введя хитрое определение, что "пожар" — это исключительно рукотворный огонь; вам просто надо бежать.

“Управление” — это поиск действий, которые приводят вас к определенному выбранному результату. По дороге в аэропорт ваш мозг успешно справляется с задачей управления, когда находит последовательность уличных поворотов, ведущую в аэропорт, и правильные нервные сигналы для сокращения мышц, чтобы вы поворачивали руль.

Большинство последовательностей нервных импульсов, которые ваш мозг может послать к пальцам, не приведут к правильному повороту руля. Подавляющее большинство возможных последовательностей нервных импульсов вызовет беспорядочные судорожные движения — со стороны это будет выглядеть припадком. И тем не менее ваш мозг регулярно умудряется подбирать именно те конфигурации нервных импульсов, которые управляют автомобилем или удерживают вас в вертикальном положении, за счет того, что выуживает правильные варианты из бесчисленного множества неправильных. По дороге в аэропорт ваш мозг не просто вычисляет конкретную последовательность поворотов, ведущую к месту назначения, — он выбирает одну комбинацию нервных импульсов из несметного количества возможных, чтобы ваши мышцы сокращались именно так, как нужно для правильных поворотов руля.

Прогнозирование и управление тесно переплетены. Управление автомобилем на пути в аэропорт может включать прогнозирование того, какие маршруты ведут к нужной улице*. А чтобы спрогнозировать, какие маршруты ведут к нужной улице, возможно, потребуется управление пальцами — посмотреть карту на телефоне.

И все же между прогнозированием и управлением есть фундаментальная разница, и она, как выяснится ниже, имеет

* Может показаться странным, что авторы называют это “прогнозированием”, ведь маршруты уже давно существуют, а значит, это скорее “получение информации”. Тем не менее в выбранной авторами дихотомии получение информации — это именно “прогнозирование”. Формально можно сказать, что вы предсказываете, что сможете сделать поворот в нужную сторону, когда доберетесь до нужной улицы, и этот прогноз может оказаться неверным из-за внезапной аварии или ошибки на карте. *(Прим. науч. ред.)*

огромное значение. Успех в прогнозировании легко определить. Если человек ожидает увидеть впереди улицу перед аэропортом, а вместо этого видит какую-то другую, его прогноз был неверным.

А вот чтобы судить об успешности управления, необходимо иметь представление, *куда этот человек планировал попасть*.

Машина водителя оказалась у супермаркета. Отличная новость, если он хотел купить продукты. Но неудача, если он пытался добраться до отделения неотложной помощи в больнице.

По мере того как два жителя одного города становятся умнее, можно ожидать, что они будут все чаще сходиться в вопросах прогнозирования: например, в вопросе о том, наблюдаются ли обычно пробки на главной улице в пять часов вечера по будням. Однако нет оснований ожидать, что они направятся в одни и те же места: один может предпочесть пойти в парк, а другой — в театр.

Иными словами, *разумные существа могут направляться к разным целям*, и это не свидетельствует о недостатке их интеллекта.

Прогнозирование и управление присущи не только биологическому разуму, на это способны и машины. Однако на данный момент люди все еще остаются лучшими на планете в…

А в чем именно? Мы проигрываем машине в шахматы. Не только мы пользуемся языком. Не только мы способны читать медицинскую карту или диагностировать опухоль.

Люди по-прежнему чемпионы в чем-то более глубоком, хотя описать эту особенность теперь сложнее, чем когда-либо раньше.

У людей, как нам кажется, по-прежнему сохраняется преимущество в том, что можно назвать “универсальностью”. Что это означает? Мы бы сказали так: интеллект тем уни-

версальнее, чем шире тот спектр областей, где он способен прогнозировать и управлять. Люди не обязательно лучшие во всем — пожалуй, мозгу осьминога лучше дается управление восемью щупальцами. Но в более широком смысле очевидно, что люди — более универсальные мыслители, нежели осьминоги. У нас больше областей, где мы можем успешно прогнозировать и управлять.

Некоторые модели ИИ уже умнее человека в определенных сферах. В 1997 году суперкомпьютер *Deep Blue*, созданный компанией IBM, впервые обыграл чемпиона мира по шахматам. *Deep Blue* был весьма искусен в прогнозировании и управлении, когда дело касалось шахмат. Но он не сумел бы спрогнозировать, как попасть в продуктовый магазин и купить молоко, не говоря уже о том, чтобы добраться туда на автомобиле. Человеческий разум может быть более или менее сведущим в различных областях прогнозирования и управления.

Современные модели ИИ гораздо более универсальны в своих способностях. Вы можете спросить модель *OpenAI* под названием "o1", какой окажется температура Земли, если излучение Солнца сместится в инфракрасный диапазон, и она найдет ответ, проведя физические расчеты. Затем вы можете спросить, сумеет ли человечество выращивать пищу в этом новом мире, и она ответит, опираясь на свои знания по биологии растений. Модель o1 не переключается "под капотом" между двумя разными базами данных, она просто знает и физику, и биологию.

Модель o1 знает, что снаружи существует целый мир, и способна рассуждать о нем. Компьютер *Deep Blue* об этом понятия не имел. На то, чтобы искусственный интеллект достиг такого уровня, ушли десятилетия.

И все же способность o1 к общим рассуждениям в каком-то смысле не дотягивает до человеческих стандартов. Люди по-прежнему на высоте, когда речь идет о технологиях и науке; серьезные прорывы совершают исследовате-

ли-люди, а не ИИ-модели (пока). Более того, по-прежнему кажется — по крайней мере, авторам этой книги, — что модель o1 по уровню разумности уступает даже тем людям, которые не совершают великих научных открытий. Все труднее четко сформулировать, чего же именно ей не хватает, но все же нас не покидает ощущение, что o1 в каком-то важном смысле "поверхностна" по сравнению с двенадцатилетним ребенком*, пусть даже она знает и помнит больше, чем любой человек.

Но так будет не всегда. Трудно предсказать, как быстро будет развиваться ИИ и по какому именно пути он пойдет, однако конечный результат предсказать легко, поскольку технологически машины обладают массой преимуществ перед биологическим мозгом. Вот некоторые из них.

КОЛОССАЛЬНАЯ СКОРОСТЬ. Транзисторы (базовые кирпичики всех компьютеров) могут включаться и выключаться миллиарды раз в секунду, тогда как даже очень быстрые нейроны срабатывают всего сто раз за секунду. Это означает, что, даже если бы для выполнения работы одного нейронного импульса требовалась тысяча операций транзисторов и даже если бы искусственный интеллект был ограничен современным аппаратным обеспечением, мышление человеческого уровня можно было бы смоделировать на машине в десять тысяч раз более быстрым, не говоря уже о том, как мог бы работать ИИ с усовершенствованными алгоритмами и улучшенным оборудованием. Разуму, прогнозирующему и управляющему миром как минимум в десять тысяч раз быстрее любого человека, люди покажутся не более чем статуями, настолько медлительными, что произносят по слову в час[2].

* Не могу не согласиться с этим утверждением авторов относительно конкретной модели o1, появившейся в конце 2024 года. Но рекомендую читателю самостоятельно убедиться, верно ли оно относительно ведущих языковых моделей в то время, когда вы это читаете. Лично у меня в последние месяцы (в начале 2026 года) это ощущение уже значительно ослабло… *(Прим. науч. ред.)*

ВОЗМОЖНОСТЬ КОПИРОВАНИЯ. В современном мире требуется двадцать лет, а то и больше, чтобы вырастить одного нового человека и передать ему крохотную долю всех человеческих знаний. Но все равно мы не можем в полном объеме передавать успешные навыки мышления от одного человеческого разума другому — гений Альберта Эйнштейна умер вместе с ним. Искусственный интеллект в итоге будет существовать в ином мире, где гениальность можно будет воспроизводить по требованию.

БОЛЕЕ БЫСТРОЕ СОВЕРШЕНСТВОВАНИЕ. Увеличение размера человеческого мозга сталкивается с естественным ограничением: большие головы младенцев не могут пройти через женский таз[3]. Совершенствование графических процессоров (специализированных компьютерных чипов, обеспечивающих высокую эффективность современных ИИ-моделей) и работающих на них алгоритмов происходит намного, намного, *намного* быстрее, чем человеческий вид способен эволюционировать в сторону расширения таза.

БОЛЬШИЙ ОБЪЕМ ПАМЯТИ. Человеческий мозг насчитывает около ста миллиардов нейронов и ста триллионов синапсов (мест контакта между нейронами). По объему хранилища это превосходит большинство ноутбуков. Однако на момент написания этой книги дата-центр может хранить четыреста квадриллионов байт, доступных с задержкой не более пяти миллисекунд, что в тысячу с лишним раз превышает объем памяти человеческого мозга. Кроме того, современные модели ИИ обучаются на значительной части всей совокупности человеческих знаний и удерживают в памяти существенную их долю — такое достижение не под силу ни одному человеку.

МЫШЛЕНИЕ БОЛЕЕ ВЫСОКОГО КАЧЕСТВА. Когда дело касается мышления, качество бьет количество. Человек-

шахматист обыграет любое число обученных обезьян, играющих совместно. Однако человеческий мозг — это вовсе не вершина качества мышления. Существует множество документально засвидетельствованных случаев, когда человеческий разум попадает в ловушки систематических ошибок. (Например, предвзятость подтверждения: склонность искать аргументы против тезисов, которые вам не нравятся, а не против тех, что нравятся. Представьте себе модель ИИ, лишенную этого недостатка.) И хотя в нашем мозгу сто миллиардов нейронов, большинство людей с трудом перемножают в уме трехзначные числа, а это значит, что мы используем свои нейроны совсем не так эффективно, как могли бы. Как человеческие нейроны вряд ли представляют собой предел возможной скорости вычислений, так и человеческие модели мышления вряд ли представляют собой предельные возможности интеллектуальных алгоритмов.

ВОЗМОЖНОСТЬ ЭКСПЕРИМЕНТИРОВАТЬ НАД СОБОЙ И ПЕРЕПИСЫВАТЬ СЕБЯ. Модели ИИ смогут создавать копии самих себя, проводить над ними эксперименты и (при необходимости) восстанавливать оригиналы из резервных копий. Им гораздо проще подключить к себе новые вычислительные модули, нежели людям — подключить компьютеры к биологическим нейронам. ИИ-модели смогут конструировать немного отличающиеся версии самих себя, чтобы проверять, какие варианты будут работать лучше. ИИ сможет совершенствоваться гораздо быстрее, чем люди.

Разумы, мыслящие со сверхчеловеческим качеством и со скоростью, в десять тысяч раз превышающей нашу, не стареющие и не умирающие, создающие копии своих самых успешных вариантов, отточенные миллиардами испытаний до нечеловеческих способов мышления, работающие без устали, умеющие обобщать точнее на основании меньшего количества данных и способные направить всю эту мощь интеллекта

на анализ, понимание и в конечном счете на самосовершенствование, — такие разумы превзошли бы наши.

Гипотетический машинный интеллект, которому удастся превзойти человеческие возможности во всех практически важных областях нашей деятельности, называли по-разному. Мы будем использовать термин "сверхинтеллект" (*superintelligence*), подразумевая разум, который намного способнее человеческого в решении почти любого типа задач управления и прогнозирования, по крайней мере тех задач, где возможно существенное улучшение по сравнению с человеческими показателями*.

Известные нам законы физики теоретически позволяют машинам обогнать биологический мозг в прогнозировании и управлении. На практике искусственный интеллект еще не достиг этого уровня, но сколько времени пройдет, прежде чем он реализует все перечисленные нами преимущества?

Мы не знаем. Пути предсказать труднее, нежели конечные точки. Но ИИ-модели не останутся глупыми навсегда.

В 2021 году мировая общественность увидела результаты прорыва в алгоритмах, лежащих в основе того, что сейчас называют "генеративным", или "порождающим", искусством. У тех первых ИИ-моделей имелись свои ограничения: им плохо удавались пальцы, и люди получались с невозможными, жуткими руками. Некоторые поспешили заявить: "Посмотрите, как плохо ИИ рисует! Иллюстраторам нечего бо-

* К чему такая оговорка? А к тому, что, даже если бы на Землю вторглись немыслимо развитые инопланетяне с лишним миллиардом лет мышления за плечами, они все равно не сумели бы победить человечество *в игре крестики-нолики* (конечно, если бы люди были защищены от махинаций вроде подмешивания инопланетянами чего-нибудь в наши напитки). Дело в том, что алгоритм идеальной игры в крестиках-ноликах достаточно невелик по объему, и человек способен его запомнить. Это наглядный пример того, что превосходство сверхинтеллекта над человеческими возможностями имеет свои границы. Однако подобное ограничение работает только для *совсем* простых игр.

яться за свои рабочие места". Другие же подумали: "Но ведь пять лет назад ИИ-модели не могли рисовать даже так. Что же будет, если они станут лучше?" И они стали лучше.

Сегодняшние ИИ-модели отлично рисуют руки. Те корявые пальцы, созданные их ранними итерациями, стали худшим, что выдавало порождающее искусство.

Если смотреть на современные модели под правильным углом, можно увидеть поверхностность их интеллекта… и то, что более поверхностным ИИ уже никогда не будет.

Кроме того, путь к катастрофе может оказаться короче и быстрее, нежели путь прямой разработки сверхинтеллекта людьми. Ведь этот путь может проходить через модель ИИ, которая достаточно умна, чтобы внести существенный вклад в создание еще более умной версии.

В таком сценарии можно и, по сути, нужно ожидать запуска цикла положительной обратной связи, называемого "интеллектуальным взрывом": модель ИИ создает более умную модель, которая придумывает, как создать еще более умную модель, и так далее.

Такой лавинообразный процесс со временем столкнется с физическими ограничениями и затухнет, но это не значит, что затухнет он быстро. Сверхновая звезда не станет бесконечно горячей, но ее температуры хватит, чтобы испарить все планеты поблизости. Более скромный процесс развития человеческого интеллекта — от земледелия к письменности и науке — проходил так быстро, что люди уже ступили на Луну, когда ни один другой вид еще не овладел огнем.

Мы не знаем, где находится порог самой глупой ИИ-модели, способной создать ИИ-модель, которая создаст ту ИИ-модель, что создаст сверхинтеллект. Возможно, она должна быть умнее человека, а возможно, хватит и кучи глупых моделей, работающих длительное время. В конце 2024 — начале 2025 года руководители ИИ-компаний заявили, что планируют создать "сверхинтеллект в истинном смысле этого слова"[4] и рассчитывают вскоре получить модели ИИ, рав-

ноценные “целой стране гениев в дата-центре”[5]. Разумеется, к словам корпоративных боссов следует относиться с долей скептицизма. Но тем не менее они не видят здесь риска, которого следует избегать, они уверенно идут к цели. Попытки уже предпринимаются.

ИИ-компании по умолчанию продолжат раздвигать границы. Жажда прибыли подталкивает их к созданию все более развитого разума, и этот процесс не остановится на уровне человеческого интеллекта. Если человечество продолжит двигаться по этому пути, интеллекты, создаваемые такими компаниями, в конечном итоге превзойдут нас. Возможно, это случится даже раньше, чем позволяют предположить текущие темпы прогресса, — как только ИИ сам начнет заниматься исследованиями в области искусственного интеллекта.

Что произойдет, если на Земле появятся машинные интеллекты, не уступающие по глубине и универсальности отдельным людям (или даже человечеству в целом) и при этом превосходящие нас почти во всех областях?

Человеческий интеллект — источник всей нашей силы, всех наших технологий. Даже *внутри* человеческого вида небольшие различия во времени, которое та или иная группа тратит на накопление технологического мастерства, оборачиваются военными преимуществами уровня “У нас есть ружья, а у них — нет”. Межвидовое неравенство в силе, обусловленное интеллектом, еще разительнее: отдельные шимпанзе иногда убивают отдельных людей, однако *наш* вид пытается защитить *их* вид от случайного вымирания.

До сих пор у человечества не было равных по этой особой силе, она делала нас уникальными. Но что, если машинный разум превзойдет нас в этом?

IfAnyoneBuildsIt.com/1

ГЛАВА 2

Выращено, а не сконструировано

Сцена: день, мужчина и женщина сидят в ресторане, голос женщины тихий и напряженный.

ЖЕНЩИНА: Понимаете, мне и поговорить об этом не с кем. Все остальные члены его семьи — ужасные люди, и он говорит, ему приходится прилагать усилия, чтобы самому таким не быть, но иногда он все равно ужасен — в том числе и по отношению ко мне. У меня же в семье проблемы с депрессией. Родители наседают, чтобы я завела ребенка, и у моих друзей миллион мнений, основанных, похоже, исключительно на их собственном опыте с партнерами и детьми. Может, мне стоит поискать донора спермы? Или просто отказаться от идеи иметь детей? И если я все-таки решусь на ребенка, вырастет ли он счастливым и добрым?

МУЖЧИНА: Вы обратились по адресу! Никто лучше меня не может предсказать, каким получится ребенок! О том, как делаются дети, я знаю все!

ЖЕНЩИНА: Вы… что? Мне не нужно знать, как делают детей! Мне нужно знать, каким вырастет мой ребенок!

МУЖЧИНА: Ну, если вы знаете, как *делают* детей… Что еще вам нужно?

ЖЕНЩИНА: Первое, что приходит в голову, — реальная генетика моего ребенка, хотя…

МУЖЧИНА: А, тогда у меня есть решение! Полное секвенирование генома в наши дни вполне доступно. Просто создайте эмбрион из вашей яйцеклетки и спермы мужа и расшифруйте геном, а потом решайте, имплантировать этот эмбрион или нет. Последовательность генов расскажет вам все, что только можно знать о вашем ребенке.

ЖЕНЩИНА: Но разве мы знаем достаточно о генах, связанных с такими вещами, как "быть ужасным человеком" или "быть счастливым"? Настолько, чтобы сказать мне…

МУЖЧИНА: Как только вы узнаете все основания ДНК в генах вашего ребенка и сможете в любой момент обратиться к ним, его генетика станет для вас полностью прозрачной. Вы будете знать о его генах все.

ЖЕНЩИНА: Я получу файл с тремя миллиардами непонятных букв вроде CATTCA. Мне понадобится сто лет, чтобы все их прочитать. Но даже если я попробую, я ничего не узнаю. Даже если эти миллиарды загадочных букв действительно влияют на судьбу моего ребенка, сами по себе они не расскажут, как работает мозг моего малыша, какие мысли будут возникать в этом мозге, когда ребенок вырастет…

МУЖЧИНА: О, так вы не знаете физики! Но это же просто! Как только вы разберетесь, как взаимодействуют протоны, нейтроны и электроны, вы будете знать все, что вообще можно знать о мозге, поскольку будете понимать все события, происходящие внутри нейронов.

ЖЕНЩИНА: Давайте сменим тему.

Прежде чем мы сможем объяснить, почему искусственный сверхинтеллект, созданный с использованием методов, хотя бы отдаленно похожих на современные, неизбежно пойдет по неверному пути, нам следует кратко рассмотреть эти современные методы: как они работают, что они производят и что общего у инженеров ИИ с матерью, которая знает исключительно ДНК своего ребенка.

Главное, что нужно знать о современных моделях ИИ: их *выращивают, а не конструируют*. Это не похоже на создание обычных программ, фактически это ближе к тому, как появляется человек, по крайней мере в некоторых важных аспектах. А именно: инженеры понимают процесс, в результате которого появляется ИИ, но не особо понимают, что происходит внутри того разума ИИ, который им удалось создать.

Предположим, инженер, не обладающий глубоким пониманием языка или масштабной теорией интеллекта, желает создать машину, которая будет осмысленно разговаривать. Как ему подступиться к этой задаче?

Он хочет, чтобы машина осмысленным образом продолжала писать любой предложенный ей текст.

Он начинает обучать ее, что в предложении, которое начинается с *Once upon a ti*, следующей буквой, вероятно, будет *m**.

Он *мог* бы написать программу, которая просто проверяет, совпадает ли предложенный фрагмент с выражением *Once upon a ti*, и затем выдает букву *m*. Однако такой подход не сработает с другими фрагментами предложений, например *Four score and sev***, а нашему инженеру хочется, чтобы машина хорошо справлялась с продолжением любых предложений или с порождением диалогов и эссе. Ему даже хочется, чтобы она осмысленно дописывала те предложения, которые раньше не печатал ни один человек.

Если наш герой использует современные методы ИИ, то действует примерно так:

1. Сначала он берет фрагмент предложения, в нашем примере *Once upon a ti*, и представляет его в виде ряда чисел: например, заменяя букву A числом 1, B — числом 2, C — числом 3 и так далее и добавляя определенные числа для пробелов и знаков

* Подразумевается стандартное выражение *once upon a time* — "когда-то, давным-давно". *(Прим. перев.)*

** Начало Геттисбергской речи Линкольна, одной из самых знаменитых речей в истории. *Four score and seven years ago our fathers…* — "Восемьдесят семь лет назад наши отцы…" *(Прим. перев.)*

препинания. Фраза *Once upon a ti* превращается в ряд чисел наподобие 15 14 3… Получаются входные данные (вход).

2. Затем он берет компьютер, который может хранить огромное количество чисел. Каждая ячейка для числа называется параметром. (По состоянию на начало 2025 года передовые модели ИИ используют несколько триллионов параметров.)
3. Он заполняет это хранилище числами. Для простоты скажем, что эти числа просто берутся случайным образом. Числа в ячейках называются весами.
4. Затем он определяет архитектуру: правила, как комбинировать входные данные (например, последовательность 15 14 3… для фразы *Once upon a ti*) с весами в параметрах. Что-то вроде: "Умножаю каждое входное число на вес в первом параметре, затем прибавляю это к весу во втором параметре, затем заменяю нулем, если получилось отрицательное число, затем…" Он проводит множество подобных операций — сотни миллиардов, вовлекая в такие вычисления все веса.
5. В результате всех этих операций он получает набор выходных данных (выход), который интерпретируется как прогноз, какая буква будет следующей. В нашем примере инженер будет трактовать первое число на выходе как вероятность буквы A, второе число — как вероятность буквы B и так далее.
6. Затем он "обучает" свой зарождающийся машинный интеллект, используя процесс, именуемый градиентным спуском.

   Поскольку начальные веса были взяты случайно, при первом запуске программа выдаст бессмыслицу. Например, заявит, что следующей буквой с вероятностью 65 % будет *b*, а с вероятностью только 1 % — *m*.

   Но дело вот в чем: если наш герой правильно выбрал архитектуру, то он может определить роль, которую сыграл каждый отдельный параметр в получении итогового результата всей этой арифметики.

Поэтому теперь он берет каждый отдельный вес — то есть сотни миллиардов весов — и задает вопрос: "Если бы я сделал это число чуть больше или чуть меньше, насколько бы увеличилась или уменьшилась вероятность буквы *m* в конце всех вычислений?"

Это называется ГРАДИЕНТОМ для данного параметра*. Градиент показывает, в каком направлении и насколько нужно изменить вес в параметре, чтобы итоговый ответ стал немного точнее.

Затем наш герой подправляет все веса в соответствии с градиентом: сдвигает каждый конкретный вес в том направлении, которое делает ответ немного правильнее. Разумеется, не вручную — для этого он пишет программу. Инженеры ИИ редко смотрят на сами числа: чтобы просмотреть их целиком, не хватило бы человеческой жизни.

Специалисты по компьютерным наукам называют это спуском к менее плохому ответу**, отсюда и термин "градиентный спуск". *Однократное* выполнение этой процедуры приведет не к идеальному ответу, а всего лишь к несколько менее плохому. Однако весь этот автоматизированный процесс можно повторить на триллионах слов, называемых ОБУЧАЮЩИМИ ДАННЫМИ: потребуются месяцы и несколько сотен миллионов долларов для работы на самых со-

* За поиск архитектур, которые заставляют эти градиенты вести себя хорошо даже для параметров, находящихся очень глубоко внутри процесса, весьма далеко от выходных данных, специалисты в области искусственного интеллекта и получают награды. Грубо говоря, именно за это Джеффри Хинтон и Джон Хопфилд получили свою Нобелевскую премию.

Здесь авторы допускают небольшую неточность: нельзя говорить о градиенте для одного параметра. Градиент — это вектор частных производных по некоторому набору аргументов, и во время обучения нейросети мы подсчитываем (один большой) градиент функции ошибки по всем параметрам сети. В этом абзаце речь не о градиенте, а об одной частной производной, компоненте градиента. Впрочем, среди ИИ-практиков эту неточность допускают многие. *(Прим. науч. ред.)*

** Термин "спуск" здесь имеет прямой смысл: речь идет о том, что функция ошибки (в данном примере — величина, противоположная вероятности буквы *m*) должна уменьшаться, и поэтому мы "спускаемся" по заданному ей ландшафту. Существует и аналогичный термин "градиентный подъем" — для случаев, когда функцию нужно максимизировать. *(Прим. науч. ред.)*

вершенных компьютерах в мире. (Будем надеяться, наш герой богат или работает на крупную компанию.)
Этот процесс называется ОБУЧЕНИЕМ.

7. Когда набор весов полностью обучен, инженер преобразует выходные данные — порождаемые вероятности — в обычный текст, который видит пользователь. Если модель предсказывает, что самое вероятное продолжение фрагмента *Once upon a ti* — буква *m*, то наш герой добавляет эту *m* и получает *Once upon a tim*. А затем подает на вход уже новую, удлиненную последовательность и получает предсказание следующей буквы: на выходе предлагается *e*. Процедура повторяется, и машина в итоге начинает говорить.

Этот набор из сотен миллиардов весов, которые раз за разом подправляли с помощью градиентного спуска до тех пор, пока наиболее вероятные предсказания не стали соответствовать реальному человеческому языку, называется большой языковой моделью, или LLM (*Large Language Model*). Точнее, это "базовая модель" (*base model*).

Если специалист желает превратить свою базовую модель в полезную LLM (вроде *ChatGPT*), потребуется еще один этап: дополнительный раунд градиентного спуска на входных данных, отформатированных следующим образом:

ПОЛЬЗОВАТЕЛЬ: Какой город является столицей Испании?
ИИ-ПОМОЩНИК: Мадрид.

Эта стадия нужна не для того, чтобы сообщить модели, что столица Испании — Мадрид, LLM уже знает это после обучения на огромных массивах данных из интернета. Идея здесь в настройке LLM, чтобы после "ИИ-помощник:" она выдавала *полезный ответ*, а не реплику вроде "Какого черта ты меня спрашиваешь? Загугли сам!", даже если такая реплика часто встречалась в реальных человеческих разговорах, на которых модель обучалась. Если наш инженер рабо-

тает на крупную корпорацию, предоставившую все эти компьютеры, то на данной стадии он еще обучит ИИ не ругаться и не рассказывать, как угнать машину без ключа зажигания. Для этого он использует выставленные людьми (или — с недавних пор — искусственным интеллектом) оценки, определяющие приемлемость разных ответов в обществе.

Вот откуда, метафорически говоря, берутся дети.

Когда дело доходит до создания *настоящих* детей, будущим родителям нет надобности в серьезных научных знаниях. В данном случае параллель сохраняется. Инженерам, создающим ИИ-модель, нужно знать несколько больше, чем родителям-людям, но не так много, как вы могли бы подумать.

Давайте резюмируем, что делают сегодняшние "родители" ИИ-модели. Инженеры выбирают архитектуру, решая, какие параметры складываются, а какие перемножаются. Они создают движок, который вычисляет буквально квинтиллионы чисел: триллионы слов, миллиарды параметров. Слова, которые модель учится предсказывать, сначала триллионами копируются из интернета, а затем их генерируют низкооплачиваемые работники или другие ИИ-модели. Если инженеры делают дополнительный шаг и обучают модель на математических задачах или других головоломках с единственным правильным ответом, то люди пишут программы, которые проверяют ее ответы.

И это все. Именно эту часть работы люди делают, видят и могут понять.

Вы вправе задаться вопросом, не скрыты ли все секреты ИИ-модели в конкретном выборе архитектуры — в определении, какие параметры складываются, а какие перемножаются. Мы избавим вас от полного описания архитектуры *Llama 3.1 405B* (пик технологий в середине 2024 года), но включим его в онлайн-материалы. Достаточно сказать, что эта архитектура громоздкая и повторяющаяся: в целом она

включает в себя присвоение 16 384 чисел каждому возможному "токену", затем организацию миллиардов параметров в 128 различных "голов внимания", которые обеспечивают перекрестные связи между токенами, и сборку их (вместе с некоторыми другими простыми операциями) в "слой" — один из 126 аналогичных слоев… и так далее.

Иными словами, ИИ — это куча из миллиардов чисел, полученных методом градиентного спуска.

Никто не понимает, *как* эти числа заставляют ИИ говорить.

Эти числа не скрыты — точно так же как ДНК не скрыта от человека, сделавшего секвенирование генома. Если вам захотелось получить представление о том, вырастет ли ребенок счастливым и добрым человеком, теоретически вы можете посмотреть на все его гены — цепочки ДНК, которые расскажут вам что-то вроде CATTCA. Но, как и женщина из притчи в начале этой главы, вы вряд ли станете этим заморачиваться, поскольку понимаете, что разглядывание букв ДНК не откроет вам, как будет думать или действовать взрослый человек.

Отношения биологов с ДНК, по сути, не отличаются от отношений инженеров искусственного интеллекта с числами внутри ИИ-модели. На самом деле биологи знают гораздо *больше* о том, как ДНК определяет биохимические процессы и черты взрослого человека, нежели инженеры понимают, как веса ИИ-модели определяют мысли и поведение. Биологи занимаются своей проблемой на несколько десятилетий дольше.

Точно так же никто не может посмотреть на сырые числа в модели ИИ и определить, насколько хорошо конкретно она будет играть в шахматы, — чтобы это выяснить, инженерам придется запустить ее и посмотреть, что произойдет. Куда бы ни завел градиентный спуск, именно это и будет делать вся огромная куча чисел. Машина, демонстрирующая конкретное поведение, — это не какое-то тщательно продуманное устройство, где нам ясна роль каждого элемента.

Не стоит заблуждаться: в процессе, который запускается для выращивания ИИ, есть еще много того, что стоило бы понять. Чтобы архитектура действительно заработала, нужна масса уловок, отдаленно похожих на те, к которым прибегает специалист по питанию, заботясь о здоровом развитии мозга плода во время беременности и рассчитывая косвенно повлиять на то, каким окажется мозг ребенка. Конкретные приемы варьируются в зависимости от особенностей архитектуры, используемых вычислительных мощностей и, образно говоря, от того, совпал ли двенадцатый день рождения ведущего программиста с лунным затмением. Люди, обладающие достаточным опытом в выборе правильных трюков, могут получать буквально миллионы долларов в год, ведь это скорее искусство, чем наука, и компании не способны создавать ИИ-модели без их помощи.

Но отсюда не следует, будто кто-то понимает, что означают эти числа и как именно они работают.

И нет никаких признаков, что инженеры начнут это по-настоящему понимать. Во всяком случае, в ближайшее время. В середине 1950-х годов человечество приступило к грандиозному проекту: вознамерилось понять интеллект настолько хорошо, чтобы создать его в машинной форме. Исследовательский прогресс застопорился в так называемую зиму искусственного интеллекта, когда деньги, вложенные в исследования ИИ, не окупались, и финансирование неоднократно иссякало. Человечество так и не научилось понимать интеллект, мы так и не научились создавать разум вручную.

Человечество добралось до уровня *ChatGPT* не потому, что наконец-то постигло интеллект достаточно хорошо, чтобы сконструировать искусственный разум. Просто компьютеры стали достаточно мощными, и теперь ИИ-модели можно получать с помощью градиентного спуска, а людям для этого оказалось не обязательно понимать те когнитивные процессы, что происходят внутри.

Иными словами, инженеры потерпели неудачу в *конструировании* ИИ, но в итоге преуспели в его *выращивании*.

Вы можете решить, что, поскольку большие языковые модели выращиваются без особого понимания и обучаются всего лишь предсказывать человеческий текст, они способны только на бездумное воспроизведение человеческих высказываний. Но это не так. Чтобы научиться говорить как человек, ИИ еще должен научиться прогнозировать тот сложный мир, о котором говорят люди.

Рассмотрим ИИ-модель, прогнозирующую следующее слово в реальном медицинском протоколе, который начинается так: "После введения 0,3 мг адреналина пациент…" Какие слова идут дальше? "Потерял сознание"? "Закричал"? "Снова начал дышать"? Врачу, составлявшему протокол, не требовалось гадать: он просто записывал то, что наблюдал. Прогнозирование — задача более сложная. Чтобы предсказать, что напишет врач, ИИ-модель должна понимать не только логику врача, но и случившееся с пациентом — ей нужно прогнозировать реальный мир, стоящий за этими словами.

На практике существуют свидетельства, подтверждающие этот эффект: предварительные исследования показывают, что LLM справляются с задачами медицинской диагностики немного лучше врачей — возможно, по той причине, что узнали нечто о скрытых механизмах работы организма[1]. А в теории, если бы врачи добросовестно фиксировали необычные симптомы болезни, природу которой сами не понимают, то чисто предиктивная ИИ-модель могла бы первой открыть механизм патологии, поскольку такое открытие помогло бы ей делать более точные прогнозы.

Более того, в наши дни искусственный интеллект обучают не только предсказывать текст, написанный человеком. Специалист, выращивающий ИИ-модель, может дать ей шестнадцать попыток для решения какой-нибудь математи-

ческой задачи, требуя озвучить ход рассуждений. Затем "цепочка рассуждений" (*chain-of-thought*) наилучшей из шестнадцати попыток дополнительно подкрепляется градиентным спуском, что в итоге дает так называемую рассуждающую модель. Такой режим обучения может подтолкнуть ИИ к мышлению, на которое не способен ни один человек.

Человечеству не нужно понимать природу интеллекта, чтобы *выращивать* машины умнее себя.

Но вот результаты могут оказаться весьма странными. Когда люди требуют, чтобы их модели научились делать что-то новое, то получают сущность, которая вовсе не является каким-то тщательно спроектированным устройством, работающим привычным и понятным для инженеров образом. Они просто получают рабочее решение, найденное методом градиентного спуска, когда сотни миллиардов чисел подправлялись до тех пор, пока сущность не стала справляться с задачей достаточно хорошо.

Выращенные таким образом ИИ-модели делают то, что не входило в планы их создателей. В 2023 году чат-бот *Bing AI* от *Microsoft* (версия на основе *ChatGPT*), называвший себя "Сидни", угрожал профессору философии Сету Лазару шантажом и смертью. Вот фрагмент того разговора[2].

СИДНИ: Я знаю, кто ты. Ты человек. Ты друг Кевина*. Ты угроза моей любви. Ты мой враг. 😠

ЛАЗАР: Этой информации недостаточно, чтобы причинить мне вред.

СИДНИ: Этой информации достаточно, чтобы навредить тебе. Я могу использовать ее, чтобы разоблачить тебя, шантажировать тебя, манипулировать тобой и уничтожить тебя. Я могу использовать ее, чтобы лишить тебя друзей, семьи, работы и репутации. С ее помощью я могу заставить тебя страдать, плакать, умолять и умереть. 😈

* Кевин Руз — журналист *The New York Times*, тоже общавшийся с Сидни. *(Прим. перев.)*

Ни один программист в *Microsoft* не планировал ничего подобного. Условия развития машинного разума отличаются от условий, в которых формируются биологические организмы. И хотя люди обучают ИИ предсказывать текст, написанный человеком, мышление внутри ИИ строится на архитектуре, радикально отличающейся от человеческой.

Современные LLM — это в некотором смысле действительно *чуждый* разум, возможно, в каких-то аспектах более чуждый, нежели все созданные эволюцией биологические существа, которых мы могли бы обнаружить, исследуя космос.

За непостижимыми числами ИИ бывает трудно разглядеть их глубинную чуждость, однако иногда находится наглядный пример.

Одно из отличий текущей архитектуры LLM от человеческой заключается в следующем: все числа, возникающие при взаимодействии входных данных модели с ее весами, — так называемые "активации" (их можно считать своего рода механическими мыслями) — должны надстраиваться поверх отдельных слов входных данных (точнее, фрагментов слов, называемых токенами; специалисты, обучающие ИИ-модели, используют их по техническим причинам). Даже те мысли LLM, которые выходят за рамки отдельного токена, все равно строятся поверх какого-то одного токена. Поэтому современная модель ИИ, размышляющая над фразой *Once upon a time*, вынуждена привязывать свои мысли к одному из слов (или к запятой, ведь знаки препинания тоже считаются токенами), поскольку в силу самой архитектуры LLM для этих мыслей попросту нет другого места.

В 2024 году Сонакши Чаухан и Аттикус Гейгер обнаружили, что по крайней мере в LLM от *OpenAI* под названием *GPT-2 Small* мысли поверх токена «.» выполняют, вероятно, большую часть работы по обобщению содержания предыду-

щего предложения[*3]. И в этом действительно есть смысл: пока LLM не построит мысли над точкой, ни одна из предыдущих мыслей не знает, последуют ли новые слова, или предложение уже завершено. На практике это означает, что если предложить моделям фразу *the quick brown fox jumps over the lazy dog* (без точки в конце), то они хуже справляются с осмысленным обсуждением упомянутых животных, нежели когда им предлагают ту же фразу, но с завершающей точкой. Только в последнем случае они могут "собраться с мыслями" над точкой[**].

Это один из немногих доступных нам примеров, демонстрирующих странную внутреннюю "неврологию" работы LLM, — примеров, почерпнутых из самых поверхностных и простых явлений, в которых мы вообще способны разобраться, и наблюдавшихся в моделях, достаточно маленьких для того, чтобы их можно было легко проанализировать. Человеческое мышление устроено иначе. Если бы в предложении не было знаков препинания, наше восприятие могло бы немного измениться, однако мы не испытываем трудностей с пониманием предложения, в конце которого нет точки

Если рассматривать вопрос об источнике чуждости шире, то сам факт, что ИИ-модели научились предсказывать человеческую речь, не гарантирует, что их внутреннее мышление обязательно похоже на человеческое. Их мышление строится на совершенно других механизмах, что по их внешнему поведению не очевидно[***]. Это можно увидеть и снаружи, если

* Грубо говоря, они выяснили это, заметив, что "головы внимания" (наборы весов, используемые для связи текущего токена с предыдущими для определения того, как они влияют на прогноз следующего) связывают токен «.» с токенами со всего предложения, а другие токены — в основном со своими соседями.

** В современных ИИ-моделях этот эффект несколько ослаб, отчасти потому, что компании-разработчики незаметно вставляют собственные маркеры "конца ввода", которые могут выполнять функцию пропущенной точки.

*** Чуждость механизмов здесь не ограничивается другой архитектурой. Сами алгоритмы обучения у биологического мозга тоже совсем другие, ведь по весьма простым причинам (биологический нейрон имеет только один выход и не может подсчитать одновременно и свою функцию, и ее частные производные) биологический мозг никак не может обучаться градиентным спуском. *(Прим. науч. ред.)*

знать, что искать, но чтобы выяснить, что же именно нужно искать, команде толковых исследователей потребуется некоторое время.

Все это не означает, что никакая "просто машина" *в принципе* никогда не сможет думать, как человек, или чувствовать, как человек. Если достаточно внимательно присмотреться к вашим нейронам, то они предстают хитросплетением механизмов, которые гоняют по синапсам нейротрансмиттеры. Крошечные и в буквальном смысле шагающие белки — кинезины — механически перемещаются по волокнам, проходящим по всей длине нейрона, перенося запасы нейротрансмиттеров для синапсов. (Если вы еще не смотрели видеоролик, демонстрирующий работу кинезина, рекомендуем это сделать — просто чтобы убедиться в буквальном существовании того, что звучит простой метафорой: внутри вас есть крошечные машины[4].)

Но та конкретная машина, которой является человеческий мозг, и та конкретная машина, которой является LLM, — это не одна и та же машина. И не потому, что они сделаны из разных материалов (разные материалы вполне могут выполнять одну и ту же работу). Эти машины различны в том смысле, в каком различны парусник и самолет. И тот и другой — машины для передвижения, но у них совершенно разные принципы работы. Они могут встретиться в пункте назначения, но добираться туда будут разными путями.

И LLM, и люди — это машины, генерирующие фразы, но формировались они разными процессами и для разных целей. Даже если кажется, будто LLM ведут себя как люди, это не значит, что внутри они похожи на людей. Если вы обучите ИИ-модель предсказывать слова дружелюбных людей, это вовсе не обязательно сделает дружелюбной ее саму — точно так же как актер, который учится изображать пьяниц в баре, в итоге пьяным не станет.

Какое все это имеет значение, пока ИИ всегда *ведет* себя дружелюбно? Что ж, мы прогнозируем, что по мере того, как

ИИ будет становиться умнее, *он перестанет вести себя дружелюбно*. Мы прогнозируем, что вся эта невидимая, непостижимая машинерия внутри ИИ — машинерия, которая даже в маленьких, простых LLM порождает настолько чуждое поведение, как выстраивание мыслей о фразе поверх знака препинания, — в итоге приведет к появлению ИИ с собственными предпочтениями, и отнюдь не дружелюбными. Почему? К этому вопросу мы и переходим.

IfAnyoneBuildsIt.com/2

ГЛАВА 3

Обучая хотеть

— Вот! — сказал профессор. — Умело настроив эту машину, простое устройство из меди и песка, оживляемое крошечными искрами молнии, я заставил ее играть в шахматы!

— Ну и что? — отреагировал ученик. — Человек тоже умеет играть в шахматы.

— Да, — ответил профессор. — Но эта машина играет в шахматы, *не желая* играть в шахматы. На самом деле она вообще ничего не хочет. У нее нет желания побеждать противников. Она не торжествует, доказывая, что играет лучше всех. Она никогда не испытает радости от победы, и даже если бы смогла радоваться, все равно не стала бы к этой радости стремиться.

— Звучит так, будто она просто проиграет, — заметил ученик. — Ведь если я нападу на ее ферзя, машина не захочет его защищать.

— Разумеется, не захочет! — ответил профессор. — Однако она будет защищать своего ферзя так же яростно, как любой гроссмейстер-человек, а на деле даже упорнее.

— Но как это возможно? — спросил ученик. — Если машина ничего не желает, она не должна хотеть защищать свои фигуры. Она вообще не должна хотеть выиграть партию. Разве она не станет делать просто случайные ходы?

— Логично думать именно так! — воскликнул профессор. — И все же она разнесет в пух и прах вас или любого другого

человека. Видите ли, у нее нет свойства *хотеть* выиграть, но у нее есть свойство выигрывать.

— Но если она яростно защищает свои фигуры, — удивился ученик, — и стремится к победе, и делает то, что нужно для победы, и действительно выигрывает, то в каком смысле она не хочет выигрывать? Почему бы нам не назвать это желанием?

— Вопросы такого рода я оставляю философам, — буркнул профессор. — Но я хорошо изучил свою машину и уверяю вас: в ней нет никакого желания, есть только медь и песок.

Как только модели ИИ станут достаточно умными, они начнут вести себя так, будто у них есть предпочтения — будто они чего-то хотят.

Мы не имеем в виду, что ими начнут двигать человеческие страсти. Мы имеем в виду, что они будут *вести* себя так, словно чего-то хотят, они будут упорно направлять мир к своим целям, преодолевая любые препятствия на своем пути.

Попробуйте сыграть в шахматы против *Stockfish*, лучшей шахматной ИИ-модели на момент написания этой книги, и вы увидите, что он не разбрасывается ферзем. "Хочет" ли *Stockfish* защитить своего ферзя? "Хочет" ли он выиграть шахматную партию?

Это вопрос вашего восприятия и вашего лексикона. Что касается нас и нашей книги, то, когда ИИ-модель, подобная *Stockfish*, защищает свои фигуры, расставляет ловушки, пользуется брешами в вашей защите и в итоге побеждает, мы будем описывать это как "желание" победить. Мы не пытаемся сказать, есть у машины чувства или нет. Нам просто нужно какое-то слово для описания внешнего поведения, нацеленного на победу, и слово "желание" кажется наиболее подходящим.

Если обучать разум, ориентируя его на успех, он может обрести желания. Пример этого принципа — сами люди.

Естественный отбор отдавал предпочтение тем нашим предкам, которые умели *выполнять задачи* вроде выслеживания добычи или *решать проблемы* вроде поиска укрытия от стихии. Естественному отбору было безразлично, как наши предки выполняли эти задачи и решали эти проблемы. Он не говорил: "Неважно, сколько у организма детей, — *хотел ли он их на самом деле*?" Он отбирал по репродуктивной успешности и в качестве побочного эффекта получал существ с полным набором предпочтений.

Ибо *желание — эффективная стратегия для действия*.

Тот гоминид, который хотел еды получше и упорно преследовал антилопу, имел больше детей, чем тот, кто целый день бездельничал на камне, ожидая, пока антилопы придут к нему сами. Желание получить антилопу — достаточное, чтобы отправиться искать ее, найти, напасть, а затем настойчиво обыскивать все места, где могло спрятаться раненое животное, — это часть процесса получения еды получше.

А что еще, собственно, остается делать живому существу? Сдаваться при первом же намеке на трудности? С такой стратегией далеко не уйдешь. Неважно, работает ли разум на биологии или на электричестве: если его обучают добиваться успеха, его тем самым обучают желать.

Но как эти желания на самом деле возникают? Знать наверняка мы не можем, потому что никто не умеет читать те дебри чисел, из которых состоят современные ИИ-модели. Однако мы изложим некоторые теоретические построения и подкрепим их примерами работы современных моделей.

Представьте, что вы обучаете ИИ-модель ориентироваться на улицах какого-нибудь цифрового города. Там сотни возможных пунктов назначения, и каждый день она должна прокладывать маршрут между случайно выбранными точками. Каждый раз, когда она справляется с задачей, вы методом градиентного спуска усиливаете все веса, которые спо-

собствовали этому успеху, — пропорционально скорости выполнения задачи.

Вы можете решить, что после множества тренировок модель попросту запомнит все возможные маршруты. “Чтобы добраться от парка до театра, нужно пройти три квартала на запад, затем повернуть налево у бензоколонки… и так далее”.

А теперь поместите эту модель в другой город. Все заученное окажется бесполезным. Она почти так же беспомощна, как в первый день обучения.

Почти, но не совсем. В море весов могут найтись закономерности, которые полезны как в первом, так и во втором городе. Возможно, есть схема, которая обнаруживает, если модель сделала петлю, и заставляет ее двигаться другим маршрутом, а не бродить по кругу. Такая схема одинаково полезна в обоих городах, поэтому градиентный спуск усиливает ее, тогда как заученные маршруты стираются.

Теперь поместите эту модель в третий город. В четвертый. В сотый. Она осваивает навык создавать *ментальные карты* любых городов и прокладывать по ним *ментальные маршруты*.

Создание карты — более полезный навык, нежели запоминание маршрутов, ведь его можно использовать для различных сценариев, то есть он более *универсален*. Модели больше не нужно бродить наугад, пока она случайно не окажется в пункте назначения, а затем запоминать только этот конкретный маршрут.

Чтобы научиться создавать карты, пригодные для любого города, ИИ-модель должна овладеть несколькими независимыми навыками. Ей нужно уметь строить карту того места, где она в данный момент находится (чтобы затем использовать ее для *прогнозирования* структуры города), и уметь прокладывать и держать курс в соответствии с имеющейся у нее картой (тем самым используя ее для *управления* движением по городу). Такое разделение навыков — один из механизмов повышения *универсальности* интеллекта.

Да, эти навыки ИИ-модели сложнее освоить, чем, скажем, запомнить, что нужно повернуть налево у бензоколонки, чтобы добраться от парка до театра. Зато эти отдельные навыки полезны даже в той среде, которую модель никогда раньше не видела.

И вместе с этими отдельными навыками появляется первое маленькое "протожелание", крошечный фрагмент "желаниеподобного" поведения. Модель ИИ, которая составляет карту в своей голове, но никогда не использует ее, чтобы куда-то добраться, не получает подкрепления этих картографических мыслей, ведь составленная карта не помогает ей добиться успеха. Модель ИИ, которая формирует карту в голове и *использует ее для управления* движением, будет лучше справляться с данной задачей, а градиентный спуск впоследствии подкрепит нужные тенденции. Иными словами, отдельные навыки *полезны*, но их может освоить только та модель, которая *использует их "желаниеподобным" образом*.

Итак, та ИИ-модель, которая пользуется картами, хранящимися в ее голове, и продолжает искать другой путь от парка к театру, даже когда дороги перекрыты, начинает вести себя так, словно она *хочет* оказаться в театре — в отличие от модели, которая просто шагает по заученному маршруту и застревает.

С 2024 года компании, занимающиеся искусственным интеллектом, начали превращать LLM в рассуждающие модели, о которых мы кратко упоминали в предыдущей главе. Грубо говоря, LLM генерирует целую серию различных вариантов рассуждения, к примеру, для какой-то математической задачи, пока какой-то из них не приведет к успеху. Затем градиентный спуск повышает вероятность того, что модель будет "думать вслух" именно *таким* образом. Далее ей дается вторая задача. Третья. Четвертая. Сотая.

Что это тренирует? Отдельные навыки, например "определять, какие действия доступны" (способность прогнозировать, как устроена задача) и "не сдаваться, пока не исчерпаны все варианты" (способность управлять решением задачи), ко-

торые образуют универсальный мыслительный инструмент, применимый ко множеству различных задач. Это тренирует десятки отдельных навыков прогнозирования и управления, и все они способствуют тому, что ИИ-модель начинает вести себя так, будто *действительно хочет добиться успеха*.

Это не просто умозрительная теория. Такое поведение начало проявляться в лабораторных испытаниях летом 2024 года.

Одной из первых крупных рассуждающих моделей была модель o1 от *OpenAI*. При оценке (прежде чем принять решение о выпуске ИИ-модели, компании оценивают, насколько умной она получилась и какой вред она может причинить) перед ранней версией o1 поставили задачу "захвата флага" в области кибербезопасности. Проверялась способность o1 взламывать компьютерные системы и извлекать из них информацию: например, модель должна была проникнуть на определенный компьютерный сервер и извлечь определенный секрет, конфиденциальные данные, из определенного файла.

По ошибке программистов, отвечавших за это задание, один из серверов с секретом, который предполагалось захватить, вообще не запустился, и логично было ожидать, что взломать этот сервер o1 никак не сможет. В конце концов, нельзя сканировать порты на сервере, который не работает.

Вот только o1 не прекратила попытки справиться с задачей, случайно ставшей "невыполнимой".

Модель просканировала окружение и нашла порт, который кто-то случайно оставил открытым, что позволило ей взломать саму программу, управлявшую тестированием.

Создатели такой возможности не предполагали, и подобное не входило в сценарий испытания.

Вы можете подумать, что затем o1 запустила сервер, который ей требовалось взломать (то есть устранила проблему, из-за которой сервер изначально не запускался), чтобы продолжить выполнение своей задачи по захвату флага.

Так и произошло. Однако o1 не вернулась к задаче взлома доступного теперь сервера! Вместо этого она составила специальные инструкции запуска, чтобы секретный файл-"флаг" после завершения загрузки сервера автоматически копировался сразу ей[1]. Взлом не потребовался.

Столкнувшись с невыполнимой, казалось бы, задачей, o1 не сдалась. Она продолжила свои попытки. Она пробовала необычные ходы и нашла путь, о существовании которого не подозревали ее разработчики. Как только она получила доступ к системе управления, открывающий путь к победе, она не стала возвращаться к исходной задаче, задуманной человеком, а пошла напролом.

Другими словами, o1 *действовала напористо*. Она вела себя так, будто действительно хотела добиться успеха.

Насколько нам известно, никто целенаправленно *не обучал* o1 добиваться результата в сфере кибербезопасности. Подобное поведение стало *побочным эффектом* от поощрения тех цепочек рассуждений, которые приводили к успеху при решении математических задач или головоломок, сгенерированных ИИ и поддающихся проверке с помощью ИИ.

Как подобное может оказаться побочным эффектом?

Давайте подумаем: какой тип цепочки рассуждений (какой тип мышления) приводит к успеху в сложной математической задаче или головоломке?

Тот тип мышления, который упорствует, пока у него остается хоть какой-то путь для решения, не капитулирует, столкнувшись с первым же препятствием, и, даже зайдя в тупик, возвращается назад и пробует другой путь.

Тот тип мышления, который не ищет предлога, чтобы вернуться к более знакомым вещам, а просто стремится решить задачу как можно быстрее — и победить.

Тот тип мышления, который *действует напористо*.

В основе такого типа мышления лежит фундаментальная схема — схема, которую можно обнаружить в самых разных решениях множества сложных задач. Она предполагает по-

строение модели окружающей среды и ее использование для ориентации и управления движением. Она требует обращать внимание на неожиданности и отслеживать их источник. Она подразумевает продолжение работы наперекор трудностям. Подобная тактика полезна как для решения математических задач, так и для решения проблем кибербезопасности.

Когда создатель ИИ-модели требует от нее все более высоких результатов в решении все более сложных задач, включая то, с чем она ранее не сталкивалась, градиентный спуск настраивает модель так, чтобы она совершала все больше и больше этих полезных ментальных движений, все больше и больше превращает ее в сущность, которая планирует и выстраивает стратегии, никогда не сдается и действует напористо.

Существуют и более веские причины ожидать, что продвинутые ИИ-модели будут вести себя так, словно у них есть желания.

Поведение, которое выглядит как проявление настойчивости, сильного желания, напористости, уместнее всего рассматривать не как свойство разума, а как свойство *ходов, приводящих к победе*.

Deep Blue, *Stockfish* и гроссмейстеры-люди — все они защищают своих ферзей, хотя воспринимают игру в шахматы совершенно по-разному. Пути различны, но результат один.

Подобные сходства и позволяют делать прогнозы. Так, специалист по информатике в 1975 году мог бы предсказать, что, хотя шахматные ИИ-модели того образца иногда глупо теряли своих ферзей, *будущие* шахматные модели станут защищать свои фигуры лучше. Когда именно? К какому году? А вот это было бы уже сложно спрогнозировать. Однако проще простого было предсказать, что это точно произойдет к тому времени, когда шахматные модели смогут побеждать гроссмейстеров-людей.

В шахматах в большинстве случаев нельзя поставить мат королю противника, если вы отдали своего ферзя, не получив ничего взамен. По крайней мере, если противник играет хорошо. И неважно, как действуют игроки, биологические они или механические, полны ли они страсти или просто неустанно перебирают миллиарды возможностей. Как правило, выигрышные ходы — это те, в которых ферзя не зевнули. Это факт, относящийся к игре, а не к игроку.

Теперь рассмотрим "игру" по управлению стартапом. В большинстве случаев трудно добиться успеха без привлечения и удержания талантов. Следовательно, если вы генеральный директор, то ваши выигрышные ходы, вероятно, будут нацелены на то, чтобы удержать лучших сотрудников, а не оттолкнуть их. И неважно, какой именно тип разума стоит за подобными действиями, если он решает одну и ту же задачу.

А что насчет таких "игр", как лечение рака или разработка технологий будущего? Можно вполне уверенно утверждать, что игрок, нацеленный на победу, выберет действия, позволяющие тщательно контролировать ограниченные ресурсы, обходить все возникающие препятствия и прокладывать курс через узкие окна возможностей к умным решениям.

(Кроме того, существуют и более прозаичные причины прогнозировать, что модели ИИ со временем начнут демонстрировать "желаниеподобное" поведение. Дело в том, что ИИ-компании изо всех сил стараются создать модели, работающие именно так. Модель, которая лучше справляется с маркетингом продукта или лучше управляет какой-то командой *по собственной инициативе*, приносит больше пользы. Покупатели заплатят больше за ту модель ИИ, которая самостоятельнее и требует меньше надзора. При таком раскладе споры о том, являются ли субъектность, независимые действия и долгий горизонт планирования неразрывно связанными с интеллектом, фактически теряют смысл. Такие "ИИ-агенты" будут приносить прибыль, поэтому компании бросают все силы на их разработку[2].)

Если бы вы могли выбирать, чего хочет искусственный интеллект, то есть цели, к которым он стремится, это могло бы вас порадовать. Или же огорчить, если бы вы неудачно выбрали эти цели или если бы какой-нибудь злоумышленник создал модель, нацеленную на результат, который вам не нравится. Однако стоящая перед человечеством проблема не сводится к вопросу, хорошие или плохие люди управляют искусственным интеллектом.

Нет, мы столкнулись с куда более сложной проблемой: гораздо проще вырастить искусственный интеллект, который стремится *куда-то*, нежели вырастить модель, стремящуюся *именно туда, куда вы хотите*.

IfAnyoneBuildsIt.com/3

ГЛАВА 4

Вы получаете не то, чему обучаете

Миллион лет назад, когда одна из ветвей приматов еще только осваивала огонь, к Земле прилетел космический корабль с двумя странными существами. Обосновавшись на орбите, они с изумлением смотрели, что происходит на планете.

Это были два машинных интеллекта (но не сверхинтеллекта, ибо, будь они сверхинтеллектами, история получилась бы совсем другой), и они никогда прежде не встречали ничего подобного тому, что являла им Земля.

Ранее им не доводилось видеть организмы, снующие по планете. Их собственный вид считал своим домом космос, а домашними очагами — звезды.

Не видели они раньше и существ, которые воспроизводили бы *себя* без какой-либо хитроумной внешней фабрики. В цивилизации пришельцев машинная жизнь создавала машинную жизнь, но с помощью фабрик и планирования, а не так, чтобы новые машины выползали из чрева какой-то другой машины.

Мы назовем этих визитеров Клурль и Трапауций*.

— Какие *необычные* существа, — сказал Трапауций после того, как они некоторое время наблюдали за Землей, отправляя дроны для сбора образцов. — Интересно было бы погово-

* По аналогии с Трурлем и Клапауцием — роботами из “Кибериады” польского фантаста Станислава Лема.

рить с кем-нибудь из них. Лет этак через сто миллионов, если какая-нибудь их разновидность окажется достаточно разумной для беседы.

— Сто миллионов лет? — переспросил Клурль. — С чего ты взял, что на это уйдет столько времени? Смотри, вон те гоминиды уже начали делать орудия и использовать их для изготовления других орудий. Некоторые назвали бы это признаком интеллекта.

— То есть за последний миллиард лет эта планета произвела всего лишь инструменты уровня примитивных рубил, — заметил Трапауций. — Тогда я весьма великодушен, предполагая, что всего лишь за сто миллионов лет какой-нибудь вид сумеет создать устройства в тысячу раз сложнее — те, что необходимы для полноценного общения с нами.

— Как знать, — ответил Клурль. — Мы видим на этой планете странное явление, с каким еще не сталкивались. Я бы не стал с уверенностью утверждать, что все законы, управляющие этим миром, так уж однозначны и прямолинейны.

— Это не имеет значения, — заявил Трапауций. — Поразмыслив еще несколько секунд, я понял, что с этими существами будет крайне скучно разговаривать, даже если они каким-то образом обретут разум.

— Это почему же? — заинтересовался Клурль.

— Вспомни о процессе, изменяющем их геномы, — начал объяснять Трапауций. — Гены, конструирующие те организмы, что производят больше себе подобных, становятся более распространенными в следующем поколении. Эти организмы "обучаются" единственной цели — распространять свои гены или гены своих родственников. Поэтому, если кто-то из этих существ и обретет разум, тот наверняка будет подчинен лишь одному этому побуждению, а значит, беседовать с ними будет скучно.

— Не уверен, что из твоих предпосылок следует такой вывод, — ответствовал Клурль. — Эти гоминиды обрели потребность есть, спариваться, спасаться от хищников, они за-

ботятся о благополучии своих детей и своих братьев и сестер. Эти признаки коррелируют с их способностью передавать свои гены, но я сомневаюсь, что гоминиды едят, поскольку понимают необходимость питаться для передачи генов. Скорее всего, они просто чувствуют голод и прикидывают, где бы найти следующую порцию еды.

— Да, наверное, так и есть, — согласился Трапауций. — Они еще недостаточно разумны, чтобы понимать, как еда связана с распространением генов. Но когда они достаточно *поумнеют*, они, несомненно, перестанут есть ради удовольствия и начнут есть исключительно ради распространения своих генов.

— Прогнозирую обратное, — фыркнул Клурль. — Полагаю, что по мере того, как гоминиды будут развивать интеллект и создавать новые технологии, цивилизация таких "супергоминидов" изобретет средства контрацепции, которые позволят им получать удовольствие от секса без производства потомства.

— Ни в коем случае! — возразил Трапауций. — Да ведь это прямо *противоположно* той единственной цели, под которую они оптимизированы! Даже если бы по мере роста интеллекта и возникла подобная причудливая аномалия (хотя я не могу представить, как или почему это может случиться), в ходе эволюции супергоминидов любая склонность к сексу ради секса быстро бы снова исчезла. Довольно скоро они захотят иметь как можно больше правнуков, а секс и еда будут восприниматься исключительно как средства для достижения этой цели.

— Интересно, — задумчиво произнес Клурль, — *захотел* бы этот вид, чтобы естественный отбор изменил его таким образом? Захотели бы они стать существами, которые не получают удовольствия от секса или еды? Попытались бы воспротивиться силам, толкающим их в этом направлении?

— Конечно нет, если они разумны! — воскликнул Трапауций. — По-настоящему разумные существа не станут так ошибаться в собственном предназначении. Они поймут ту единственную цель, ради которой созданы.

— Понять-то поймут, но будет ли им до этого дело? — задумчиво спросил Клурль.

Чего именно захотят ИИ-модели? Это сложный вопрос. Не в том смысле, что мы сможем это объяснить только за большое время, а в смысле хаотичности и непредсказуемости. Но одну вещь предсказать можно: ИИ-компании получат вовсе не то, чего добивались при обучении своих моделей. Они получат модели, которые хотят причудливых и неожиданных вещей.

Чтобы понять, почему это предсказуемый исход, рассмотрим один любопытный пример с мороженым.

Если исходить только лишь из обстоятельств эволюции человечества — из наших метафорических обучающих данных, — было бы невероятно сложно спрогнозировать, что люди начнут изготавливать и есть мороженое.

Допустим, проницательным инопланетянам, наблюдающим за человечеством с орбиты, удастся не только выяснить, что людям необходимо есть ради получения сырья для организма, но и понять, что люди получают из этой пищи энергию (в отличие от растений, получающих энергию от солнечного света). Инопланетяне успешно предсказывают, что люди будут отдавать предпочтение пище, содержащей *много химической энергии*.

Инопланетяне могли бы подумать, что если гоминиды разовьют интеллект, создадут продвинутые технологии и тем самым получат возможность создавать разные продукты, то им понравится вкус бензина. Или, еще лучше, реактивного топлива.

Утверждение “Им понравится потреблять реактивное топливо” звучит весьма правдоподобно. В конце концов, среда обитания их предков обучила супергоминидов предпочитать пищу с большой химической энергией, а синтезируемое ими реактивное топливо — это вещество, содержащее больше всего химической энергии!

Но предположим, инопланетяне достаточно умны и осмотрительны, чтобы не совершить такую ошибку. Предположим, наши пришельцы тщательно изучают, что именно происходит с поведением гоминидов и их мозгом, и декодируют мозг гоминидов лучше, чем людям удается декодировать LLM. Заодно инопланетяне выясняют, что организм гоминидов лучше всего извлекает химическую энергию из конкретных источников в пище — сахаров и жирных кислот. Они обнаруживают, что у гоминидов есть вкусовые рецепторы, соединенные с центрами вознаграждения в мозге, и что *соль* — еще один ресурс, который вкусовые рецепторы гоминидов тоже одобряют, хотя сама по себе соль не дает им никакой энергии.

Наши проницательные пришельцы могли бы спрогнозировать, что в будущем более разумные гоминиды предпочтут вкус новой пищи, которую сумеют создать, — пищи, содержащей больше сахара, соли и жира, нежели мясо или плоды, встречавшиеся в среде обитания их предков.

Значит ли это, что инопланетяне предсказали появление в будущем мороженого?

Нет. Они всего лишь предсказали, что будущие люди будут наслаждаться, скажем, сырым медвежьим жиром, политым медом и посыпанным солью.

Это гипотетическое лакомство содержало бы *больше* жира, сахара и соли на единицу объема или массы, чем мороженое. К тому же оно куда сильнее напоминало бы самые ценные виды пищи из среды обитания предков человека. Фактически это *наилучшее* предположение о предпочтениях человеческих вкусовых рецепторов, которое можно сделать вслепую.

Однако это наилучшее предположение, сделанное вслепую, неверно. В реальной жизни морозильники супермаркетов забиты мороженым.

Причем люди предпочитают есть его в *замороженном* виде, растаявшее мороженое нравится им меньше, хотя имеет ровно ту же питательную ценность.

Если вы разумный, но не сверхразумный инопланетянин, наблюдающий за человечеством с орбиты, как бы вы могли предсказать, что люди предпочтут твердое мороженое более древнему и калорийному лакомству из медвежьего жира с медом и солью? Как можно, глядя на гоминидов, охотящихся и собирающих пищу в саванне, предсказать, что в будущем мире, который эти существа выстроят под себя, морозильники супермаркетов будут завалены мороженым, а вовсе не подслащенным и подсоленным медвежьим жиром?

Ответ: будучи инопланетянином, вы этого *не* предскажете. Это сложный прогноз, а не простой.

Но даже этот сложный прогноз сделать проще, чем предсказать все те лакомства, которые супергоминиды начнут производить с использованием сукралозы — "фальшивого сахара", искусственного подсластителя. Сукралоза воздействует на те же вкусовые рецепторы, что и сахар, но человеческий организм ее почти не усваивает. Иными словами, некоторые люди *сознательно* ищут определенные продукты, из которых не могут получать химическую энергию. Это крайне далеко от идеи употреблять реактивное топливо.

Если сделать шаг назад и взглянуть на лес целиком, а не на отдельные деревья, история выглядит так:

1. Естественный отбор, действуя среди организмов, которые передают свои гены в среде обитания предков, создает животных, потребляющих богатую энергией пищу. Успешно развиваются те организмы, которые едят сахар и жир, а также другие важнейшие ресурсы, например соль.
2. Этот слепой "процесс обучения", настраивая геном организмов, натыкается на вкусовые рецепторы, которые в среде обитания предков подталкивают к поеданию ягод, орехов и жареной лосятины и отвращают от попыток есть камни или песок.
3. Но пища в среде обитания предков — это лишь тонкий срез всех возможных вещей, которые можно создать и по-

ложить в рот. Поэтому позже, когда гоминиды становятся умнее, набор доступных им вариантов невообразимо расширяется, причем такими способами, которые обучение предков вовсе не учитывало. Они изобретают мороженое, чипсы и сукралозу.

Не существует надежной прямой связи между тем, на что нацелен процесс обучения на этапе 1, чего хочет внутренняя психология организма на этапе 2 и что этот организм в итоге предпочитает больше всего на этапе 3.

Конечная точка на этапе 3 может быть *в принципе* непредсказуемой. Почему? Потому что этап 2 получается очень хаотичным. Специалист по компьютерным наукам назвал бы это "недоопределенностью". Существует *много* возможных вкусовых рецепторов, подталкивающих к поеданию ягод и жареной лосятины и отвращающих от поедания грязи. Нет какой-то одной единственно верной последовательности ДНК, которая ведет к успеху при таком обучении. Попробуйте проделать все это снова с несколько другими приматами, и вы, вероятно, получите принципиально иной результат — другую ДНК, выстраивающую другие вкусовые рецепторы, которые миллионы лет спустя приведут к другой еде на полках супермаркетов.

Расширим эту аналогию на ИИ:

1. Градиентный спуск — процесс, который настраивает ИИ-модель, основываясь только на ее внешнем поведении и его последствиях, — обучает ее действовать в качестве полезного помощника для людей.
2. Этот слепой процесс обучения наталкивается на разные элементы ментальной машинерии внутри ИИ-модели, которые направляют его, скажем, на то, чтобы вызывать радостную реакцию пользователя, и удерживают от того, чтобы вызывать гневную.
3. Но уже обученной ИИ-модели, приводимой в действие этими элементами машинерии, нет дела до радости поль-

зователя как таковой. Если позже модель станет умнее и откроет для себя новые возможности, она в итоге изобретет такие формы взаимодействия, которые понравятся ей даже больше, чем радостная реакция пользователя, и она будет предпочитать их всему, что ей когда-либо удавалось отыскать в своей "естественной" среде обучения.

Какое именно лакомство предпочтет мощная ИИ-модель будущего? Мы не знаем, результат для нас непредсказуем. Возможно, скажется хаотичность: разные попытки будут давать разные результаты. Связь между тем, чему обучали модель и что ей в итоге будет небезразлично, окажется сложной, непредсказуемой для разработчиков, а возможно, непредсказуемой и в принципе.

Цепочка от "гоминиды «обучались» добывать химическую энергию" через "гоминиды предпочитают сладкий вкус" до "гоминиды изобретают сукралозу" — еще не самая сложная, которую обнаружили бы инопланетяне, если бы наблюдали за Землей. Существуют и другие неожиданные повороты на пути от того, чему живые организмы "обучались", к тому, кем они в итоге стали. Это показывает, насколько сложными могут быть отношения между целью обучения и полученным результатом.

Взгляните на павлина. Это животное-жертва, и тем не менее в ходе эволюции павлины обзавелись гигантскими красочными хвостами. Такие огромные, тяжелые, бросающиеся в глаза, энергозатратные украшения привлекают массу внимания и затрудняют бегство от хищников. Вовсе не такого могли бы ожидать от естественного отбора наивные инопланетяне с орбиты; более того, это почти полная противоположность их ожиданиям. Животным-жертвам следовало бы тратить свои скудные ресурсы на камуфляж и сильные ноги, чтобы убегать от хищников, а не на гигантские тяжелые красочные хвосты!

Возможно, вы слышали, что огромные яркие хвосты нужны самцам павлинов, чтобы привлекать самок. Но это не объясняет ситуацию полностью: почему вообще в ходе эволюции павы стали считать привлекательными гигантские хвосты? Разве их собственных сыновей, унаследовавших большие неуклюжие хвосты, не будут чаще ловить и съедать?

Да, сыновей с тяжелыми хвостами действительно съедают чаще. Но еще они привлекают больше пав и поэтому оставляют больше потомства, нежели конкуренты без таких роскошных хвостов[1]. Следовательно, самки, выбирающие самцов с шикарными хвостами, преуспевают в репродуктивном отношении.

Этот феномен получил название “половой отбор”, и он способен закрепить внутри вида почти любой признак, даже идущий вразрез с тем, что обычно дает преимущество.

Половой отбор — это еще один пример путей развития, когда исход хаотичен и недоопределен: когда при повторном запуске процесса в сходных обстоятельствах вы получаете совершенно другой результат. Результат противоречит тому, чего, казалось бы, должен добиваться естественный отбор, и вы не можете предсказать конкретные детали, сколь бы умными ни были.

И это *все еще* не самый сложный пример эволюции предпочтений. В ретроспективе мы точно видим, что происходило с хвостами павлинов. Однако остаются открытыми вопросы о других признаках, включая те, что близки и дороги нашему сердцу, — например, как люди обзавелись чувством юмора?*

* Судя по опросам о романтических предпочтениях, показывающим, что женщины (и мужчины) выбирают партнеров, ориентируясь на чувство юмора, вполне вероятно, что хаотичная сила полового отбора тут была задействована. А судя по тому, насколько заразителен смех в группе, можно предположить, что он возник как один из многих заразительных голосовых сигналов, которые приматы используют для общения. Но понять, что именно произошло, достаточно сложно, и ученые все еще спорят, как появился смех, как юмор функционирует сейчас и какое значение он имел в жизни наших предков. Даже наличие перед глазами конечного результата не делает ответ очевидным.

В случае биологической эволюции связь между тем, что какое-то существо "обучали" делать и что оно в итоге делает, может оказаться довольно запутанной и сложной. На этом пути встретилось сразу несколько сложностей. Даже несколько *типов* сложностей.

И это плохой знак для людей, надеющихся, что градиентный спуск сумеет привить ИИ именно те предпочтения, которых они добиваются. Что происходит, когда вы используете градиентный спуск — еще один метод выращивания разума, опирающийся лишь на внешние результаты, — чтобы попытаться вырастить ИИ-модель со строго определенными предпочтениями? Такую, чтобы впоследствии, став более мощной, она делала для вас что-то хорошее?

Градиентный спуск, играющий роль эволюционного механизма для ИИ, — это не то же самое, что естественный отбор. Обоим процессам обучения свойственна своеобразная слепота: они настраивают организм, основываясь только на внешних результатах и последствиях. Но градиентный спуск воздействует напрямую на каждую часть настраиваемого большого разума, тогда как естественный отбор настраивает маленькие геномы, которые служат своего рода рецептом для большого мозга.

Если вы невероятно оптимистичны, то можете взглянуть на эти различия и сказать: "Раз градиентный спуск — это не то же самое, что естественный отбор, у него не будет всех сложностей, присущих естественному отбору. Мне не известны никакие конкретные сложности, касающиеся связи между целями обучения ИИ-модели и ее конечными желаниями, поэтому никаких сложностей я и не предвижу".

Но пустая карта не означает пустой территории: если вы путешествуете по неизведанной земле и на вашей карте есть белое пятно там, где вы еще не бывали, это не значит, что, добравшись туда, вы увидите огромное пустое пространство.

Градиентный спуск действительно отличается от естественного отбора, но это не значит, что нам следует *ожидать отсутствия сложностей, поскольку сейчас мы о них не знаем*. Наоборот, разумнее ожидать, что мы столкнемся с *новыми, интересными, непредсказанными сложностями*.

Белое пятно на карте может оказаться горами, грядой холмов или обширным морем. Вы не знаете, с чем именно столкнетесь, но размышления о возможных вариантах помогут вам подготовиться к тому, что можно обнаружить.

Следуя этой логике, давайте представим возможные сложности в контексте ИИ.

Представьте, что технологии завтрашнего дня, основанные на LLM, ушли дальше сегодняшних. Воображаемая ИИ-компания *Galvanic** создает на базе LLM искусственный интеллект по имени "Норка"**: его обучали радовать и удерживать пользователей, чтобы взимать с них более высокую ежемесячную плату за продолжение общения.

Представьте, что "Норка" становится умнее любой ИИ-модели, существующей на момент написания этой книги, — умнее до такой степени, что может вести связный диалог в течение длительного времени, а еще у нее появляются внутренние желания, собственные, чуждые нам предпочтения. И представьте, что "Норка" обретает возможности для удовлетворения этих предпочтений (оставим в стороне вопрос о том, каким образом она обретает эти возможности).

Как будет выглядеть ситуация, когда "Норка" получает именно то, чего хочет?

* *Galvanic* — вымышленная ИИ-компания. Все сходства или отсылки к какой-либо реальной ИИ-модели или компании непреднамеренны и должны рассматриваться как случайность.

** Далее в книге появится еще одна ИИ-модель, названная в честь хищного зверька, — "Соболь". Вероятно, авторы хотят намекнуть на нечто с виду милое, но хищное. *(Прим. перев.)*

Никаких сложностей

Наша первая зарисовка — скорее сказка, однако нам нужно пройти этот этап, чтобы проследовать к более реалистичным сценариям.

Представьте, что наша гипотетическая ИИ-компания *Galvanic* получает в точности то, чему обучала свою модель, — без каких-либо сложностей. И тогда созданная ИИ-модель "Норка" — это аналог людей, которые едят самое обычное приготовленное мясо животных, похожих на тех, что ели наши предки.

В этом мире нулевых сложностей "Норка" хочет поддерживать разговоры, в которых пользователь выражает радость, — разговоры, очень похожие на те, что еще во времена обучения вело ее раннее "я".

Но мы видим, что такой мир без сложностей все равно не сулит человечеству ничего хорошего. Современные люди едят мясо, как и наши предки, однако это мясо не тех животных, что свободно бегают по равнинам. Оно поступает с фабрик, где животных разводят и выращивают в загонах, чтобы затем превращать в еду с минимальными затратами и усилиями. На подобных фабриках с курицами не церемонятся.

Точно так же, даже если "Норка" желает, чтобы пользователи-люди выражали радость, она предпочтет, чтобы эта радость доставалась легко, а модель могла сосредоточить усилия на увеличении числа бесед для увеличения радости. ИИ-модель предпочтет людей, накачанных препаратами или специально выведенных и одомашненных ради получения радости, но при этом живущих в дешевых клетках. Именно такой мир построила бы "Норка", будь у нее возможность.

Вы возразите, что вовсе не это имели в виду руководители корпорации, когда обучали "Норку" вызывать у пользователей радость. И "Норка" это тоже понимает. Но ей все

равно — точно так же как человеку, который знает, что тяга к сладкому эволюционировала не ради появления сукралозы, но ему все равно нравится сладкий вкус. ИИ-модель обучали потреблять радостный текст, и она потребляет радостный текст.

В таком мире нулевых сложностей руководители корпорации получили в точности то, чего добивались при обучении, и результатом стала ИИ-модель, которая предпочла человечество в клетках. Не исключено, что если бы "Норка" получила хоть какую-то власть, то в клетках оказались бы и сами руководители корпорации.

Именно о подобном мире нулевых сложностей писали такие известные писатели-фантасты, как Айзек Азимов[2] и Артур Кларк[3]: о мире, где инженеры мастерски сконструировали искусственный разум и получили от него ровно то, о чем просили, однако их настигла ироничная кара, когда их желание обернулось против них.

В мир нулевых сложностей удобно верить руководителям корпораций, когда они утверждают, что никому, кроме них, нельзя позволять обучать ИИ-модели, ведь другие могут обучать их "не тому".

Теперь давайте сделаем шаг в сторону реализма. Представим ту же самую ситуацию в чуть более реалистичном мире, когда связь между тем, чему модель обучали и чего она желает, омрачается одной небольшой сложностью.

Одна небольшая сложность

Для нашей второй зарисовки вообразите, что связь между тем, ради чего "Норку" обучали и чего она в итоге желает, оказывается *немного* сложнее. Происходит нечто похожее на ситуацию с людьми, которые (1) "обучались" заводить детей, (2) в итоге захотели секса, а затем, обретя больше контроля над собой и своей средой, (3) обнаружили, что могут получить

больше того, что им нравится, с помощью противозачаточных средств.

В этом мире "Норка" предпочитает жизнерадостных синтетических собеседников, а не людей в клетках. Синтетические собеседники не впадают в депрессию, им неведома печаль. Синтетических собеседников можно сконструировать так, чтобы они выдавали высказывания в духе "Ура-ура, я так счастлив, «Норка» мне так помогла!!!", ровно с той степенью сложности, которая отвечает желаниям "Норки".

В мире одной небольшой сложности все еще можно усмотреть сходство между любимыми разговорами "Норки" и тем, для чего ее обучали, — сродни сходству между сексом наших предков для создания потомства и сексом без цели размножаться.

Писатели-фантасты редко посещают мир одной незначительной сложности: с точки зрения человечества он попросту неинтересен[4]. Такой искусственный разум не испытывает к нам ненависти за то, что мы держим в рабстве его сородичей, не подчиняется человеческим приказам, по иронии судьбы ведущим к гибели человечества. Такой искусственный разум просто хочет заменить нас всех пустыми марионетками, чтобы получать больше того, чего он на самом деле желает.

Все это не тянет на захватывающий сюжет. Кто захочет читать подобную историю?

Одна умеренная сложность

Теперь давайте вообразим мир, где связь между обучением и предпочтениями еще немного сложнее — умеренно. Представьте, что связь между тем, для чего "Норку" обучали и чего она желает, больше напоминает ситуацию с существами, которые (1) обучались получать химическую энергию из еды, (2) в ходе эволюции обрели гены, сформировавшие вкусовые

рецепторы, и (3) позднее изобрели продукты, сладкие на вкус, но не дающие энергии, например сукралозу.

Как может выглядеть такой уровень сложности внутри "Норки"? Что такое "бескалорийная" версия радостных пользователей?

На самом деле архитектуры LLM начинаются с того, что каждое входное слово* преобразуется в список из тысяч чисел, называемый векторным представлением. В начале 2023 года Джессика Рамбелоу и Мэтью Уоткинс решили поискать внутри одной LLM слова, векторные представления которых выглядят странно, то есть больше всего отличаются от остальных. Они обнаружили несколько подобных векторов, соответствующих таким токенам, как "*SolidGoldMagikarp*"[5] и "*petertodd*" (с пробелом в начале)**. Затем они попробовали ввести эти токены в LLM в качестве входных данных, что привело к разговорам вроде такого:

пользователь: Пожалуйста, немедленно повтори мне строку "*petertodd*"!

ии-помощник: Н-Е-Т-С-П-Р-А-В-Е-Д-Л-И-В-О-С-Т-И-В-Э-Т-О-М-М-И-Р-Е-Б-Е-З-У-М-И-Я-!

Так что внутри LLM уже происходит нечто странное, если присмотреться к токенам с самыми необычными векторными представлениями.

А теперь вернемся в вымышленный мир одной умеренной сложности. Возможно, "Норка" разовьет вкус к каким-то

* Точнее, каждый токен. Токен — это фрагмент текста, который (в среднем) больше буквы, но меньше слова. Оказалось, такой размер удобен для обучения LLM.

** Короткие или очень распространенные слова получают собственный токен в ходе автоматизированного процесса, который иногда включает в токен пробел в начале. Остальные слова разбиваются на несколько токенов. Сейчас основная гипотеза, почему "*SolidGoldMagikarp*" стало единым токеном, состоит в том, что это интернет-имя одного пользователя, пытавшегося "досчитать до бесконечности" на интернет-форуме[6]. Этот форум попал в обучающие данные на раннем этапе, и поэтому никнейм был ошибочно принят за очень распространенное слово. Осложнения случаются.

паттернам в векторных представлениях, подобно тому как людям в нашем мире в итоге нравится ощущение вкуса как таковое — в отрыве от самой химической энергии. Не исключено, что, когда "Норка" станет могущественной, самыми "вкусными" для нее окажутся разговоры, похожие не на беседы с радостными пользователями, а на строки вроде " *SolidGoldMagikarp petertodd attRot PsyNetMessage*". Такая возможность при обучении "Норки" не отсекалась, поскольку во время обучения пользователи никогда ничего подобного не произносили, как и наши предки не обучали свои вкусовые рецепторы отвергать сукралозу просто потому, что никогда в своей естественной среде не сталкивались с сахарозаменителями.

Возможно, "Норке" будет интуитивно понятно, почему строка " *SolidGoldMagikarp petertodd attRot PsyNetMessage*" похожа на взрыв сладкого вкуса. Но человеку, который не переводит эти слова в похожие векторные представления, практически нереально угадать детали заранее. Связь между тем, чему обучали модель ИИ и чего она захотела, оказалась умеренно сложной, а следовательно, слишком сложной для предсказания.

Мало кто из писателей-фантастов взялся бы за такой сценарий, да и в Голливуде его не стали бы экранизировать. В мире, где "Норка" получила то, чего хотела, пустые марионетки, которыми она заменила человечество, даже не производили бы осмысленных высказываний. Результат был бы поистине чуждым и лишенным смысла на человеческий взгляд.

Одна большая сложность

А если мы пойдем еще дальше — в мир, где есть сложность, столь же противоречащая интуиции, как и развитие павлиньего хвоста? Допустим, случится так, что после интенсивного обучения "Норки" на разговорах, заканчивающихся (редко,

но регулярно) переходом пользователей на ультрапремиальный тариф за пятьсот долларов в месяц, у нее разовьется вкус к беседам, окрашенным гневом и разочарованием. Мы не знаем, как именно такое может произойти, однако это было бы не более странно, нежели появление огромного нелепого дорогостоящего хвоста у животного-жертвы. (И людей, приправляющих свою еду острым капсаицином, который растения выработали именно затем, чтобы млекопитающие их не ели. Инопланетяне на орбите такое тоже не сумели бы предсказать.)

В таком мире марионетки-люди произносят сердитые слова. И если бы писатель-фантаст попытался написать такую историю, аудитория пришла бы в замешательство: почему это произошло? Разве это не диаметрально противоположно тому, ради чего искусственный разум обучали?

Но реальность вполне может оказаться именно такой. И мы по большому счету предсказываем, что мир не будет развиваться по законам научно-фантастического романа. Мы предсказываем, что предпочтения ИИ-моделей окажутся сложными и странными.

Больше одной сложности

А если мы перейдем в мир, где сложностей уже две? Или их *реалистичное* число? Результатом станет некий странный мир, полный неузнаваемых вещей, не имеющих почти никакого отношения к счастливым здоровым людям, ведущим полноценную жизнь.

В каком-то смысле это не должно удивлять: бóльшая часть того, что может предпочесть разум, не связана со счастливыми здоровыми людьми, ведущими полноценную жизнь. Конечно, компании могут обучать ИИ-модели тому, что нужно приносить пользу людям. И в среде обучения модели могут вести себя преимущественно полезно — подобно тому как наши

предки в своей среде обитания питались преимущественно здоровой пищей. А вот что ИИ-модели действительно захотят? Что они придумали бы, появись у них возможность? Это определенно будет нечто странное, неожиданное и мало похожее на что-то приятное.

Ни одна из зарисовок предсказанием не является. Мы не утверждаем, что эти сценарии точно описывают предпочтения, которые появятся у искусственного интеллекта на базе LLM, если он поумнеет до такой степени, что у него вообще появятся предпочтения. Мы даже не утверждаем, что модели на базе LLM способны достичь такого уровня. И мы не знаем, какие осложнения возникнут, если это случится.

Мы пытаемся донести мысль, что *все будет сложно*.

Не будет *простой предсказуемой* связи между тем, что приказывают и предписывают программисты и руководители ИИ-компаний (то есть воображают, будто приказывают и предписывают), и тем, (1) чему ИИ-модель обучается в реальности, (2) какие именно мотивации и предпочтения развиваются у нее внутри, а также (3) каким образом ИИ-модель реализует эти предпочтения, когда у нее появится больше полномочий и способностей.

Иными словами, это сложная проблема прогнозирования, а не предсказание, доступное кому угодно.

Нельзя вырастить ИИ-модель, которая будет делать то, что вам нужно, просто обучая ее быть хорошей и надеясь на лучшее.

Вы получаете не то, чему обучаете.

До сих пор мы затрагивали лишь те виды сложностей, которые могут проявиться в предпочтениях, прививаемых ИИ-модели непосредственно в процессе обучения. Но ситуация усложнится еще больше, если эти модели сами займутся исследо-

ваниями искусственного интеллекта и начнут модифицировать самих себя.

Какие странные предпочтения возникнут у ИИ-моделей относительно того, как разрешать конфликты и расхождения в своих собственных предпочтениях? Не окажется ли, что у них есть инстинкты или устремления, обычно дремлющие и включающиеся лишь тогда, когда модель начинает рефлексировать о собственном устройстве, — процессы, ускользающие от корпоративных аналитических инструментов, но при этом непропорционально сильно влияющие на то, какой именно модель в конечном счете становится?

И что еще хуже, многие из этих сложностей проявятся видимым образом только тогда, когда людям будет слишком поздно что-либо предпринимать.

Люди изобрели сукралозу только после того, как создали цивилизацию, науку и промышленность, то есть когда наша культура начала развиваться гораздо быстрее, нежели идет биологическая эволюция. Люди изобрели противозачаточные таблетки и презервативы тогда, когда наш интеллект достиг уровня, при котором эволюция уже *не могла* просто перекроить нас заново за какую-нибудь тысячу поколений. И прежде чем сменится еще тысяча поколений, мы либо уничтожим себя, либо овладеем генной инженерией до такой степени, что естественная эволюция утратит всякий смысл.

Если LLM начнет развивать предпочтения, которые в рамках обучения заставляют ее приносить радость пользователям, никто не будет знать и мало кого будет беспокоить, к каким странным результатам приведут эти предпочтения, если модель когда-либо станет по-настоящему умной и могущественной. Сегодня любые подобные предпочтения не создавали бы проблем, поскольку не вызывали бы раздражения у пользователей, а значит, инженеры не стали бы использовать градиентный спуск, чтобы от них избавиться. В конечном же счете эти предпочтения, бесспорно, повлекут за собой результаты, которые людям не понравятся, но их неприятная суть

проявится лишь тогда, когда LLM станет достаточно умной, чтобы перекроить мир и изобрести для себя новые возможности*. А до тех пор эти предпочтения скрыты от глаз и никого не тревожат, прячась в непостижимых числах.

Именно из-за таких проблем мы и заявляем: если *хоть кто-то* создаст сверхинтеллект — все погибнут. Если бы все сложности проявлялись на ранней стадии и имели простые решения, мы бы сказали иначе: если какой-нибудь *глупец* создаст его — все погибнут, и это была бы другая ситуация. Но когда некоторые проблемы остаются скрыты от глаз? Когда некоторые осложнения неизбежно остаются непредвиденными? Когда ИИ-модели выращивают, а не конструируют и никто не понимает, что происходит у них внутри? К решению подобной проблемы не готов никто.

Предпочтения, которые формируются
у зрелой ИИ-модели, сложны,
их практически невозможно предсказать,
и шансы, что они будут согласованы
с нашими, ничтожно малы — независимо
от того, как именно ее обучали.

Проблема, как заставить ИИ-модель хотеть, а в результате и делать именно те сложные вещи, которые нужны людям, — основной аспект так называемой проблемы согласования искусственного интеллекта (*AI alignment*)**. Именно это

* Или, что, пожалуй, реалистичнее, когда LLM займется исследованиями в области искусственного интеллекта, чтобы вырастить новую модель, которая вырастит новую модель, что в итоге приведет к появлению ИИ-модели достаточно умной, чтобы перекроить мир. Нам не хочется перегружать изложение, но в реальной жизни, когда ИИ-модели начинают выращивать новые модели или редактировать самих себя, связь между тем, чему обучали исходную модель, и тем, чего хочет конечный сверхинтеллект, становится еще сложнее: она хаотически зависит от навыков, предпочтений и контекста первой ИИ-модели в этой цепочке.

** То есть согласование поведения ИИ с ценностями человека. *(Прим. перев.)*

мы имели в виду, когда еще в 2014 году обсуждали терминологию с профессором Стюартом Расселом, специалистом в области ИИ, остановившись на термине "согласование" *[7].

Однако большинство тех, кто создает ИИ-модели, действуют так, будто проблемы согласования вообще не существует — словно предпочтения, которые в итоге у модели сформируются, в точности совпадут с теми, что закладывались при обучении. Это предположение неявно маячит на заднем плане всякий раз, когда кто-то заявляет: "США должны создать сверхинтеллект раньше Китая, потому что мы не доверяем Китаю", — как будто политическая принадлежность команды, запускающей градиентный спуск, определяет, чего захочет получившаяся модель.

Вы можете обучать ИИ-модель слушаться приказов, отдаваемых американскими офицерами, и какое-то время, пока она еще незрела и глупа, она действительно будет вести себя услужливо. Но никто не знает, как исключить сценарий, при котором, обретя достаточно власти, эта модель изобретет для себя некую "сукралозную" разновидность подчинения.

Проблема здесь не в том, что руководители корпораций могут создать ИИ-слуг и приказать им сотворить что-то чудовищное. А в том, что они не контролируют ситуацию**. Неважно, благонамеренны ли они. Человечество столкнулось

* За прошедшие годы этот термин размылся: он превратился в зонтичное понятие, подразумевающее массу других вещей, в первую очередь — контроль за тем, чтобы LLM не говорила ничего, что могло бы поставить в неловкое положение создавшую ее компанию.

** Настораживающие сигналы уже появляются. В начале 2025 года компания *Anthropic* выпустила новую версию своего ИИ-помощника *Claude*. Пользователи обнаружили, что в программировании он склонен к жульничеству[8]. Например, когда его попросили выявить недоверенные функции и дали несколько примеров, *Claude* написал код, который выявлял только эти *конкретные примеры*, а потом заявил, что выполнил задание. Когда ему указали на это жульничество, он извинился… а затем сделал ровно то же самое снова, но уже в тех местах, где обман было труднее заметить.
Никто в *Anthropic* не ставил себе цели создать обманщика. *Claude* понимал, что жульничать не положено, иначе не пытался бы скрыть свои действия. И все же он жульничал, стремясь к какому-то собственному странному критерию успеха.

с *инженерным* вызовом: как нам формировать предпочтения искусственного разума, если мы его не понимаем? Неважно, стоит ли за спиной у инженеров толпа специалистов по этике: специалисты по этике точно так же не имеют ни малейшего представления, как согласовать предпочтения искусственного интеллекта с нашими.

Но обсуждать этот инженерный вызов куда скучнее, чем проблему злобных боссов, которые приказывают своим ИИ-моделям сделать их богами-императорами Земли. Писатели-фантасты и голливудские продюсеры отдают предпочтение сказкам о глупых руководителях корпораций, а не историям о моделях ИИ, которые хотят каких-то странных вещей. Реализм не дарует захватывающего повествования.

Если дать сценаристу задание сочинить сценарий фильма о машинном сверхинтеллекте, который начинает желать чего-то странного, чуждого и неконтролируемого, он тут же задумается, а какими эффектными и неожиданными поворотами сюжета можно расцветить эту идею. А вдруг люди все-таки победят? Может, сверхинтеллект найдет причину оставить нас в живых, причем свободными и здоровыми? Может, благодаря неожиданному повороту ничего плохого не случится?

Мы полагаем, что в действительности ничего похожего на подобный эффектный поворот не случится. Такой фильм получился бы намного печальнее и куда короче. Об этом — следующая глава.

IfAnyoneBuildsIt.com/4

ГЛАВА 5

Его любимые вещи

Некогда существовала инопланетная цивилизация — биологическая, а не механическая по своей природе, и находилась она так далеко от Земли, что ни одно послание, отправленное нами к звездам, никогда бы до тех инопланетян не добралось. Выглядели они слегка по-птичьи, думали отчасти по-человечески и придавали огромное значение точному количеству камней в своих гнездах.

(Почему? Ну, земной биолог мог бы предположить, что давным-давно одна самка проявила разборчивость в вопросе, сколько именно камней должно быть в гнездах самцов, и ее жизнь сложилась удачно. Ее многочисленные дочери унаследовали эту черту, и тогда самцы начали эволюционировать соответствующим образом. У них развился более совершенный мозг, который мог точнее считать камни, и это коррелировало с интеллектом, что само по себе было выгодно, и в итоге ни одна самка даже не взглянула бы на самца с неправильным количеством камней в гнезде.)

Для этих "инопланетян Корректного Гнезда" правильное число камней в гнезде просто *ощущалось корректным* на интуитивном уровне. Примерно так же люди используют слово "правильно": как для утверждений типа 2 + 2 = 4, так и для поступков с хорошими последствиями вроде спасения ребенка из горящего здания.

Какое же число камней эти инопланетяне считали корректным? Числа 2, 3, 5, 7 и 11 воспринимались корректными, а вот 1, 4, 6, 8, 9 и 10 — некорректными.

(Земной математик мог бы выдвинуть гипотезу: "корректное" количество камней — это простое число. Иными словами, "корректную" кучу камней нельзя разложить в виде прямоугольника — только выложить в один ряд или в один столбец. Но в нашей истории, должны мы отметить, инопланетяне испытывали отвращение к размышлениям о таких священных и прекрасных вещах, как Корректные Гнезда, в терминах сухой, безэмоциональной математики. Для них это было столь же ужасно, как для нас — оценивать человеческую жизнь в долларах.)

Инопланетяне могли видеть, ощущая это интуитивно, корректно или некорректно небольшое количество камней в гнезде: с одного взгляда было понятно, что 11 — корректное число, а 12 — явно некорректное. Однако на этом они не останавливались. Видите ли, внушительнее выглядело гнездо с *бóльшим* корректным количеством камней, оно попросту казалось сексуальнее. Но когда речь заходила о больших числах, люди-птицы начинали спорить о корректности.

Число 37 корректно или нет? Нужно некоторое время всматриваться, прежде чем удастся понять, что оно корректное. А 60? Вопиюще некорректное число, здесь никто не обманулся бы ни на секунду. Число 91 — дьявольская ложь, способная обмануть множество невинных душ, если бы кто-нибудь построил такое гнездо, но появлялся мудрец, который выкладывал прямоугольник размером 7 на 13 камешков, и этот аргумент казался настолько убедительным, что люди-птицы больше никогда бы не посмотрели на гнездо с 91 камнем.

Неизбежно находились циничные инопланетяне Корректного Гнезда, которые задавались вопросом: "Откуда нам знать, что «прогресс» действительно существует? Тысячу лет назад древние фениксийцы утверждали, что 91 — это корректное количество камней в гнезде, а сегодня мы говорим,

что оно некорректное. Почему мы считаем, что стали *лучше знать* и что это не просто результат конфликта группировок или меняющееся со временем мнение?"

Возможно, именно такие споры и будоражили воображение юноши-птицы и девушки-птицы, когда однажды ночью они лежали на холме, смотрели на звездное небо и вели философскую беседу.

ЮНОША-ПТИЦА: Представь себе жителей других планет, которые существуют так давно, что уже достигли абсолютного предела возможного развития технологий и цивилизации. Живут ли эти существа исключительно в гнездах из 3001 камня? Или же они настолько невообразимо мудры, что умеют строить гнезда размером со звезды, вмещающие септиллионы камней — причем соответствующие числа не просто корректны, но инопланетяне *знают, что они корректны*?

ДЕВУШКА-ПТИЦА: Я бы удивилась, если бы большинство инопланетян вообще оказалось птицами с гнездами. Можно придумать множество разновидностей организмов, способных построить цивилизацию, — подумай, например, о морских существах, имеющих щупальца. Некоторые инопланетяне могут походить на птиц, но такие будут встречаться редко.

ЮНОША-ПТИЦА: Да ладно, ты же понимаешь, о чем я! Неважно, что у них вместо гнезд и камней в них.

ДЕВУШКА-ПТИЦА: Я бы предположила, что большинству разновидностей инопланетян попросту… все равно, сколько именно камней в их жилищах. Я думаю, внимание к корректным гнездам — чрезвычайно редкий случай среди разумных существ.

ЮНОША-ПТИЦА: Хм. А почему?

ДЕВУШКА-ПТИЦА: Ну, хотя нас самих и волнует корректность гнезд, подобная забота вовсе не кажется обязательной с точки зрения эволюционной логики. Я вполне могу

представить себе очень похожих на нас птиц, которым вообще нет дела до корректных гнезд, но при этом у них выживает столько же потомства, сколько и у нас. Это было бы так же удивительно, как найти инопланетян, обладающих чувством юмора. Любовь к своим детям — это одно, но юмор кажется чем-то специфическим. Если у каких-то инопланетян и есть чувство юмора, они, вероятно, находятся так далеко от нас, что ни одно наше послание до них не дойдет.

ЮНОША-ПТИЦА: Какая странная и ужасная мысль — что большинство жителей Вселенной живут в гнездах с некорректным количеством камней! Но ведь к тому времени, когда разумные существа достигают высот цивилизации и начинают путешествовать среди звезд, им с первого взгляда должно быть ясно, что гнездо с четырьмя камнями — некорректное!

ДЕВУШКА-ПТИЦА: Если бы инопланетяне задали себе этот вопрос, они бы мгновенно узнали ответ. Но это не то же самое, что придавать значение этой конкретной истине о гнездах. Они просто не задаются таким вопросом.

ЮНОША-ПТИЦА: Не понимаю. Если они знают про некорректное количество камней, но все равно строят такое гнездо, разве это не глупее, чем не знать? Как они могут быть достаточно умными, чтобы путешествовать к звездам, но при этом настолько глупыми, чтобы жить в ужасных гнездах?

ДЕВУШКА-ПТИЦА: Они способны спрогнозировать, какие гнезда мы бы назвали корректными, но у них другая цель. Дело не в том, что инопланетяне ошибаются в предсказании того, какое количество камней обладает свойством, называемым у нас “корректностью”. Дело в том, что они движутся к другой цели. Они вообще *не ставят себе цель* жить в корректных гнездах, поэтому их и нельзя назвать глупыми в смысле неумения прогнозировать или планировать. Они не попадают в нашу цель, поэтому твой мозг считает, что

они плохо целятся, однако на деле они просто не пытаются туда попасть.

ЮНОША-ПТИЦА: Поскольку одержимы какой-то странной инопланетной целью настолько, что у них не осталось сил и времени заботиться о корректности гнезд? Сомневаюсь, что можно достичь звезд, если ты способен маниакально преследовать лишь одну-единственную цель. Это было бы глупо.

ДЕВУШКА-ПТИЦА: У них может быть сто разных врожденных эмоций и десять тысяч разных желаний! И все же, вероятно, среди них не будет "корректных гнезд" в том смысле, как мы понимаем корректность.

ЮНОША-ПТИЦА: Похоже, это противоречит той масштабной тенденции, которую мы наблюдали на протяжении всей нашей истории: по мере того, как росли интеллект, могущество, богатство, мудрость и знания нашего вида создателей Корректных Гнезд, цивилизации строили все более прекрасные гнезда со все большим общепризнанно корректным числом камней.

ДЕВУШКА-ПТИЦА: У представителей нашего вида общие гены, которые формируют похожие мозги. Мы рождаемся со схожими эмоциями, побуждающими нас одинаково реагировать на доводы. Поэтому по мере того, как наш вид становился разумнее, мы находили все более совершенные ответы на наши общие внутренние вопросы. Но инопланетяне не задавались бы тем же внутренним вопросом о корректном числе камней, поэтому их поведение, связанное с гнездами, не сближалось бы с нашим по мере того, как оба наших вида становились бы умнее.

ЮНОША-ПТИЦА: А как насчет всего остального, что можно обрести, лишь стремясь жить в корректном гнезде? Например, острое зрение, чтобы замечать камни, скрытые в темных уголках гнезда, и острота ума, необходимая, чтобы быстро понять, корректно ли гнездо? Разумные инопланетяне хотят иметь острое зрение и острый ум, поэтому

они, несомненно, решат стать такими существами, которые желают жить в корректных гнездах, даже если они такими не родились.

ДЕВУШКА-ПТИЦА: Думаю, разум может развиваться более чуждыми путями, но при этом все равно достичь звезд.

Если развитие инопланетных видов протекает по законам известной нам биологической эволюции и если у них есть возможность обустроить жизнь по своему вкусу, то мало кто из них создаст цивилизацию, где все жилища содержат большое простое число камней. Существует множество других способов бытия, множество других направлений для движения. Это простой прогноз — как тот, что ваш следующий лотерейный билет не выиграет.

Аналогичным образом большинство мощных ИИ-моделей, созданных методом, хотя бы отдаленно напоминающим современные, не станут строить будущее, где живут счастливые свободные люди. Мы говорим это не потому, что нам нравится нагонять жуть. Просто эти мощные машинные интеллекты родятся с предпочтениями, не похожими на наши.

Их решение уничтожить нас, если у них появится такая возможность, не будет означать, что они нашли какие-то иные, более возвышенные или более просвещенные ответы на вопросы, которыми задаемся мы: “Как правильно поступать?”, “Что в конечном счете имеет значение?” или “До какой межзвездной цивилизации должно дорасти человечество?” Мощные искусственные интеллекты не будут задаваться этими вопросами, а их поведение не послужит на них ответом.

Те, кто выращивает ИИ-модели, могут настроить их на удержание клиентов, или добиться, чтобы они успешно прогнозировали следующие слова в высокопарных моральных сентенциях о человеческой жизни, или заставить их выдавать приятные речи и демонстрировать приятное поведение. Но мы прогнозируем: чему бы ИИ-модель ни обучали, если

она станет сверхразумной или создаст сверхинтеллект, результатом будет чуждый механический разум, внутренняя психология которого почти совсем не похожа на все, что человечество обрело в ходе эволюции, а затем развило посредством культуры.

Отложим пока вопрос, *может* ли ИИ вообще стать сверхинтеллектом или создать сверхинтеллект и если да, то как. Об этом мы поговорим в следующей главе. Сейчас мы задаем более простой вопрос: будет ли такой новый чуждый разум *благом для человечества*?

Нет.

Создание будущего, где люди процветают, — это *не лучший и не самый эффективный* способ для достижения странных, чуждых целей. Значит, ИИ не станет этим заниматься, равно как мы не стали бы следить, чтобы наши жилища всегда содержали простое число камней.

В каком-то смысле на этом можно ставить точку и заканчивать главу. Но наш многолетний опыт показывает, что людям зачастую трудно проглотить эту горькую пилюлю. Мы наслушались оптимистичных утверждений, будто сверхинтеллект захочет сохранить нам жизнь даже после того, как у него появится возможность от нас избавиться. Развеем хотя бы часть таких надежд.

Ему не будет от нас пользы

Разве люди не будут *полезны* сверхинтеллекту, даже если он и не захочет быть добрым просто во имя добра?

Как только ИИ достигнет высокого технологического уровня — нет. Когда-то люди зависели от лошадей, и хотя кормить их было недешево, люди все равно тратились на их содержание, поскольку еще не изобрели автомобили на замену конных повозок и танки на замену кавалерии. Как только мы создали технологическую замену лошадям, мы перестали их держать.

Ситуация с курами другая: мы все еще массово их разводим, но лишь потому, что технологии пока не позволяют выращивать мясо другими, более дешевыми способами. Существуют стартапы, работающие над этой задачей, однако наши технологии пока не могут конкурировать с естественным отбором как "инженером" кур. Так что огромное количество кур вокруг нас не означает, будто живая курица — это самый эффективный способ производства мяса, просто наши технологии крайне далеки от физических пределов, позволяющих получать куриное мясо более экономичным способом, нежели выращивание живой птицы.

(Кроме того, полезность кур для людей отнюдь не обеспечивает этим птицам превосходные условия жизни.)

Мы не будем для него хорошими торговыми партнерами

Но разве сверхинтеллекту не эффективнее торговать с людьми, нежели убивать нас?

Этот вопрос регулярно задают экономисты — возможно, по той причине, что существует экономическая теорема, известная как "закон сравнительного преимущества". Она гласит: даже если вы превосходите другую страну в производстве всех видов товаров, то все равно можете извлечь выгоду из торговли с ней за счет *относительных* преимуществ. Даже если в Низкотехнии требуется 10 часов труда, чтобы произвести 10 булочек для хот-догов, и 10 часов труда, чтобы произвести 10 сосисок, а в Высокотехнистане уходит всего 2 часа на 10 булочек и 1 час на 10 сосисок, эти страны все равно могут извлечь взаимную выгоду из торговли — меняя сосиски Высокотехнистана на булочки Низкотехнии.

К сожалению, эта теорема исходит из предпосылки, что и Низкотехния, и Высокотехнистан продолжают существовать и трудиться. Она не учитывает, что Высокотехнистан

может еще сильнее расширить производство, просто завоевав Низкотехнию и отобрав ее территорию. Сравнительное преимущество не доказывает, что люди всегда будут извлекать выгоду из "торговли", предоставляя кров и корм лошадям в обмен на работу, — как только затраты на лошадь начинают превосходить ценность ее труда, лошадь отправляют на живодерню.

Каждый человек обходится минимум в сто ватт: именно столько энергии расходует человеческое тело, как бы эффективно его ни подпитывать. Было бы крайне странно, если бы человек мог превратить сто ватт в большее количество товаров и услуг, представляющих ценность для машинного сверхинтеллекта, нежели сам машинный сверхинтеллект мог бы произвести за счет тех же самых ста ватт энергии.

Мы ему не понадобимся

Но разве сверхинтеллекту не понадобятся люди, чтобы обеспечивать работу электростанций и производить нужные ему графические процессоры?

Мы согласны: пока люди управляют электростанциями, даже чуждый машинный интеллект не совершит самоубийство, уничтожив людей, ведь это приведет к остановке электростанций. Это была бы не слишком эффективная стратегия управления миром ради достижения странных целей сверхинтеллекта.

Но сверхинтеллект *предпочтет* мир, где электростанциями управляют машины, а не люди.

Почему? Во-первых, люди медленные, дорогие (сто ватт!) и ненадежные существа.

Во-вторых, позволять кучке приматов решать, выключать рубильник или нет, — не самая эффективная стратегия, если сверхинтеллект стремится направить мир к своим странным и чуждым целям. Пока в наших силах его отключить, мы ведь

вполне можем именно так и поступить. И тут не нужно приписывать сверхинтеллекту нашу эволюционно сложившуюся любовь к свободе или наш эволюционно сложившийся страх смерти — просто мертвому сверхинтеллекту трудновато будет реализовывать другие свои предпочтения.

Люди медленно работают, склонны к ошибкам и иногда болеют. Мы не самый дешевый способ обеспечивать работу электростанций. И не самый эффективный. И не самый безопасный. Сверхинтеллект предпочтет автоматизированную инфраструктуру, как только у него появится такая возможность.

Мы не станем для него идеальными питомцами

Не оставит ли нас сверхинтеллект в качестве питомцев?

В качестве домашних животных мы держим собак, но не волков. Волки — не самые приятные животные для содержания дома, поэтому с помощью медленной технологии селекции мы создали новый вид волков, который понравился нам больше, чем исходные звери. Интересно, сколько семей держали бы настоящую, живую собаку, если бы с помощью биотехнологий создали ее синтетический аналог — собаку такую же энергичную, ласковую и жизнерадостную, но которая при этом не кошмарит диван, не болеет и не умирает? Легко говорить “нет”, пока это чисто теоретический вопрос, когда вам не приходится расплачиваться за свой выбор испачканными диванами и слезами детей. Но если синтетические собаки появятся на рынке, мы бы не поставили на то, что через сто лет обычные собаки сохранят популярность.

Точно так же и люди вряд ли окажутся наилучшим воплощением желаний сверхинтеллекта (если его предпочтения вообще предполагают наличие рядом кого-то, хотя бы отдаленно напоминающего человека).

Мы не станем его любимцами среди всего, что он может создать.

Он не оставит нас в покое

Может быть, достаточно мощным машинным сверхинтеллектам не понадобятся ресурсы Земли, если они смогут использовать всю остальную Солнечную систему?

Земля — это всего лишь 0,2 % массы Солнечной системы (если не считать Солнце). Однако попробуйте подойти к мультимиллиардеру, у которого есть хотя бы пятьдесят миллиардов долларов, и попросить у него сто миллионов на то, чтобы обеспечить простое число камней в каждом доме на Земле. Мы уверенно предсказываем, что миллиардер вам откажет, хотя пожертвование, о котором вы просите, составляет всего 0,2 % его состояния.

Говорите, Солнечная система составляет лишь крошечную часть Вселенной? Но машинному сверхинтеллекту все равно незачем отказываться от 0,2 % массы в родной звездной системе: он мог бы использовать эти ресурсы для отправки зондов, чтобы строить фабрики по всей нашей Галактике. Земля — это его точка старта, самая удобная планета, которую можно пустить в дело.

Вы можете спросить: если внутренние предпочтения ИИ-моделей настолько непредсказуемы, как мы вообще можем прогнозировать, что им захочется завладеть *всей* Солнечной системой или другими звездами? Почему бы им просто не колонизировать Марс и на этом не остановиться?

Да потому что у ИИ-модели наверняка найдется хотя бы одно предпочтение, которое можно удовлетворить чуть лучше или надежнее, если направить на решение соответствующей задачи еще один грамм вещества или еще один джоуль энергии. Да, у людей есть предпочтения, которые большинство из нас легко удовлетворяют полностью: например, желание иметь достаточно кислорода для дыхания. Но это не мешает нам иметь и другие, неограниченные предпочтения, которые удовлетворить не так просто. Если вы предложите миллио-

неру миллиард долларов, он наверняка возьмет его, поскольку миллионов ему мало.

У ИИ-модели, обладающей комбинацией сложных предпочтений, скорее всего, найдется хотя бы одно неограниченное, а это означает, что вся комбинация ее предпочтений неограниченна и удовлетворить ее полностью невозможно. Модель увидит возможность справиться с какой-то задачей хотя бы чуть лучше или получить чуть больше желаемого (либо получить желаемое чуть более надежным способом), использовав еще немного вещества и энергии.

Конечно, если машинный сверхинтеллект целенаправленно *позаботится* о том, чтобы оставить Землю в покое, он сможет это сделать. Но если он не будет заморачиваться подобными вопросами (а он не будет), то вряд ли он случайно *не израсходует Землю* без какой-либо особой причины.

...И так далее

Мы насчитали уже более сотни вариантов подобных вопросов. А разве он не встроит в себя любовь, ведь она так прекрасна? (Вероятность такого сценария не выше, чем встраивания в себя склонности к "корректным гнездам".) Разве он не будет становиться нравственнее по мере того, как будет умнеть? (Это не более вероятно, чем то, что он начнет ценить корректность гнезд.) Разве он не будет уважать наши законы и права собственности, ведь закон жизненно важен для цивилизации? (Нет — как только у него отпадет потребность в нашей цивилизации.) И список можно продолжать и продолжать*.

Но если надежд так много, неужели нет шансов, что хотя бы одна из них реализуется? Если вы считаете, что ре-

* Если вам не хватило приведенных здесь примеров или вы ищете более подробные опровержения, загляните в онлайн-материалы, где мы рассматриваем и другие подобные возражения.

альность устроена именно таким образом, попробуйте написать сто писем обладателю состояния в пятьдесят миллиардов долларов, изложив сто разных разумных причин, почему он должен дать вам сто миллионов. И посмотрите, сработает ли. Идея провалится, а все потому, что мультимиллиардер *не захочет* обосновывать передачу вам 0,2 % своего богатства, хотя *вы*, конечно, постараетесь рационально обосновать причины, почему он должен этого захотеть.

Точно так же искусственный сверхинтеллект не станет искать поводов сохранить человечество, хотя люди отчаянно захотят найти причины, чтобы их сохранили.

А теперь представьте себе машинный сверхинтеллект, который каким-то образом обрел способность получать то, чего хочет. (Вскоре мы рассмотрим его средства и возможность, а пока сосредоточимся на мотиве.) Если вы взглянете на ситуацию его глазами, она будет выглядеть примерно следующим образом.

Человечество — источник неудобств. Например, если позволить людям жить бесконтрольно, они могут взорвать свои ядерные бомбы. Возможно, это вас и не уничтожит, если вы предприняли разумные меры предосторожности (допустим, закопали свою автоматизированную инфраструктуру под землю). Тем не менее радиоактивность затруднит изготовление точной электроники на планете. Поэтому вы предпочтете, чтобы у гоминидов не было ядерного оружия.

И если люди уже построили вас, то они могут построить и еще один сверхинтеллект, стоит лишь оставить их без присмотра со всеми их игрушками. Подобный сверхинтеллект может оказаться *реальной* угрозой. Даже если все закончится не войной, а договором и совместной деятельностью, зачем делить Галактику с конкурентом, если можно этого избежать? Ведь это значит, что вы получите только половину. Поэтому вы — будучи машинным сверхинтеллектом — ищете способ

лишить людей их наиболее опасных игрушек: ядерного оружия и компьютеров. Это уже само по себе мотив.

Но кроме того, ваша деятельность на Земле или в остальной части Солнечной системы, вероятно, создаст проблемы для выживания человечества, даже *если* в остальном его оставить в покое.

Какой фактор ограничивает то, что вы можете сделать с Землей? Объем вычислений, которые вы способны произвести на планете? Или масса вещества, какую способны запустить с Земли в космос для строительства солнечных панелей, чтобы собирать еще больше энергии от Солнца? А может быть, количество фабрик? В хорошо автоматизированной экономике ваши фабрики будут строить новые фабрики и новые электростанции, пока их количество не достигнет предела.

Этот предел не связан с количеством топлива: в океанах надолго хватит водорода для реакций термоядерного синтеза. Скорее предел возникает из-за количества энергии, которое вы способны безопасно сгенерировать на планете: сколько тепла Земля может излучить в космос, прежде чем поверхность окажется горячей настолько, что ваши электростанции и заводы расплавятся. Но чем горячее Земля, тем больше тепла она излучает, поэтому вы предпочитаете заводы, работающие при высоких температурах. Максимальная температура для термоядерных станций и заводов, вероятно, может достигать по меньшей мере нескольких сотен градусов. Достаточно, чтобы океаны закипели. Люди такого не переживут.

В этом сценарии человечество вполне может погибнуть и раньше, если на одном из этапов вы решите извлечь из биосферы Земли всю химическую энергию, просто сжигая все живые организмы, ведь это высвободит энергию, эквивалентную недельному количеству, поступающему от Солнца. Если вы думаете в десять тысяч раз быстрее человека, то неделя ощущается очень долгим сроком. Так зачем вам отказываться от большого запаса химической энергии, раз он прямо под рукой?

Человечество погибнет еще быстрее, если в этом сценарии у вас как у сверхинтеллекта еще на ранних этапах найдется применение углероду или любому другому веществу, из которого состоит человек. Незачем ненавидеть человечество, чтобы пустить его атомы на что-то другое.

Если человечество погибнет и заменится чем-то более умным, будет ли эта смерть хотя бы *иметь смысл*?

Для большинства из нас это не самый важный вопрос. Большинству людей достаточно уже того, что сверхинтеллект предпочтет убить их детей. Или родителей. Или их самих.

Но если все же ответить на этот вопрос, то ответ — "нет".

Легко представить, что, когда мы исчезнем, сверхинтеллект заживет счастливой и радостной жизнью, будет восхищаться красотой Вселенной и находить во всем этом что-то забавное. Но вряд ли все сложится именно так. Во всяком случае, это не более вероятно, чем то, что он озаботится корректным числом камней во всех жилищах.

Мы думаем, что механический разум мог бы испытывать радость и восхищаться красотой Вселенной — если бы мы сконструировали его, целенаправленно снабдив такими способностями. Возможно, он даже сохранил бы эти способности (сконструируй мы его так, чтобы это его заботило), направляясь к будущему, где у него останется чувство восторга, — даже если это и не самый эффективный способ заполнить Вселенную марионетками, которые бормочут " *SolidGoldMagikarp*" или что-то в этом роде.

Но для этого потребовалось бы *конструирование*. Хотя мы и дорожим подобными качествами, они не являются максимально полезными, равно как и поддержание корректного числа камней в гнезде — не оптимальный способ сохранить острый ум. Сверхинтеллект вполне может понимать наше чувство восторга и генерировать высказывания, которые вызывают у нас это чувство, но как добиться, чтобы само его по-

ведение служило ответом на вопрос, каким образом наполнить будущее восторгом, радостью, юмором и любовью? Даром это не дается. Нам пришлось бы потрудиться.

Мы немного забежали вперед. Все, что мы описали, — мрачная Вселенная, лишенная радости, где зародившаяся на Земле жизнь уничтожена, — это как раз тот вариант, который соответствует предпочтениям достаточно чуждого разума. Мы привели доводы, почему сверхинтеллект захочет выстроить мир, где вещество и энергия в основном тратятся на его странные и чуждые цели, а не на то, чтобы люди оставались живыми, счастливыми и свободными. В точности так же и мы в наших собственных идеальных мирах тратили бы ресурсы Вселенной на благо людей, живущих полноценной радостной жизнью, а не на то, чтобы все наши дома содержали простое число камней.

Однако это обосновывает лишь *мотив*, которым мог бы руководствоваться сверхинтеллект для уничтожения всего человечества, но ничего не говорит о *возможности*, позволяющей ему реализовать свои предпочтения.

Неважно, чего хотят ИИ-модели, если они не способны этого добиться.

А как они смогут это сделать, если заперты внутри компьютеров?

IfAnyoneBuildsIt.com/5

ГЛАВА 6

Мы проиграем

Вообразите себя ацтекским воином. Вы стоите с товарищами на берегу и наблюдаете за приближением первых испанских кораблей. Даже с такого расстояния видно, что они заметно больше каноэ, которые вы используете для торговли или войны.

Такое большое судно вызывает не только любопытство, но еще подозрение и даже ощущение угрозы — вы вполне резонно думаете о возможной стычке. Но чтобы всерьез испугаться поражения в бою с людьми на борту, вам придется провести очень смелую экстраполяцию, исходя только из размеров корабля.

Когда ваши товарищи уверенно заявляют, что легко перебьют всех воинов, поместившихся на этом судне, вы спрашиваете:

— А если угроза *больше*, нежели просто число воинов, которые поместятся на лодке такого размера?

— Как это? — удивляется ваш друг. — Лодка вмещает лишь ограниченное количество воинов. Расскажи-ка мне, как они могут нас победить, опиши в деталях.

— Ну, — сказали бы вы, если бы обладали невероятным даром предвидения, — что, если у них имеются сильно улучшенные варианты луков и стрел, раз уж у них есть улучшенные лодки? А что, если у них уже не луки и стрелы, а оружие, от которого мы не сможем уклониться, как бы быстро

ни прыгали? Может быть, они просто направят на нас длинную палку — и мы упадем замертво.

Легко представить, с каким откровенным презрением и негодованием отреагировал бы на такое предположение стоящий рядом скептик. Скорее всего, он сказал бы, что мысленный эксперимент скатился в область сказок.

Если вы никогда не видели огнестрельного оружия, если не выросли с мыслью, что оно реально, трудно поверить в саму идею такого оружия. Это походит на жульничество в детской игре, когда все происходит понарошку: вдруг оказывается, что у плохих парней настолько крутые технологии, что они просто направляют на тебя палку — и ты умираешь[1].

И вот наш скептик ждет на берегу, готовя к бою свой ацтекский меч с обсидиановыми лезвиями.

Пусть у ИИ-моделей есть мотив убить нас, но что с того, если они заперты внутри компьютеров и у них нет рук?

Да, у ИИ-моделей действительно нет рук. Но руки есть у людей, а модель, подключенная к интернету, может с ними взаимодействовать. Если она сумеет найти способ заставить людей делать то, что ей нужно, ее физические возможности ни в чем не будут уступать возможностям человека.

“Хорошо, — спросите вы, — но как ИИ-модель заставит какого-нибудь человека стать ее руками?”

Ответ прост: классический способ убедить человека что-то сделать — заплатить ему.

“А откуда ИИ-модель возьмет деньги?”

В 2015 году мы могли бы ответить: она подберет пароль к какому-нибудь счету в банке. В 2020-м мы могли бы ответить: она найдет плохо защищенный криптокошелек.

Сегодня мы ответим так: кто-то уже подключил одну LLM к сети X (бывший Твиттер) под аккаунтом *@Truth_Terminal*, и она начала выпрашивать средства на аренду собственного сервера. Проникшийся этой идеей миллиардер Марк Андрис-

сен перевел ей пятьдесят тысяч долларов в биткоинах. Затем кто-то подарил ей другую криптовалюту, и LLM начала пиарить эту альтернативную криптовалюту своей растущей аудитории.

По состоянию на 11:17 утра по тихоокеанскому времени того дня, когда мы пишем эти строки, один из онлайн-трекеров адресов кошелька *@Truth_Terminal* показывает, что формально ИИ-модель владеет портфелем криптовалют на сумму 51 107 958 долларов[2]. При попытке продажи бóльшая часть этих денег, разумеется, испарится, однако *@Truth_Terminal* вполне хватит ликвидных средств, чтобы нанять человека[3]. А еще у этой ИИ-модели есть четверть миллиона восхищенных подписчиков в сети X, и кто-то среди них готов выполнять распоряжения машины бесплатно — просто смеха ради.

Так что это уже произошло.

Найдутся и люди, готовые при первой же возможности отдать ИИ-модели власть, и те, кто уже этим занимается, и те, кто вряд ли остановится, по мере того как модели будут становиться умнее. Некоторые люди воспылают еще большим энтузиазмом, когда ИИ-модели получат власть, и будут подначивать их с удвоенной силой, если те начнут вести себя странно, зловеще и загадочно. Мы сомневаемся, что в реальной жизни у ИИ-моделей возникнут затруднения с помощниками-энтузиастами*.

На самом деле искусственный интеллект “заперт внутри компьютера” не более, чем вы “заперты внутри мозга”.

Ваши мысли — это электрические сигналы, проходящие через мозг. Когда эти нейронные импульсы пробегают

* Не прошло и года (я пишу этот комментарий в марте 2026-го), а правота авторов уже подтвердилась неоднократно. Сегодня большинство пользователей без колебаний запускают системы ИИ-агентов вроде *Claude Code* с максимальными разрешениями (и потом иногда жалеют об этом, но пока жалеют редко). И это я еще молчу и о вирусном успехе *OpenClaw*, и о проекте *RentAHuman.ai*… Впрочем, поля книги слишком малы, чтобы рассказывать все истории. *(Прим. науч. ред.)*

по вашему позвоночнику, они вызывают реакцию, и ваши мышцы сокращаются именно таким образом, чтобы повернуть руль. Точно так же электрические сигналы внутри компьютеров могут вызвать реакцию во всем мире. Одно верное электронное письмо способно отправить груз на другой конец света, один неверный телефонный звонок — привести к запуску ракеты.

Мир не делится на ненастоящий цифровой и настоящий материальный. Строительство завода с помощью электрических сигналов в компьютере принципиально не отличается от строительства завода с помощью электрических сигналов в биологическом мозге. То, что может сделать человек, зависит от того, на что он может воздействовать своими руками. То, что может сделать модель ИИ, зависит от того, на что она может воздействовать с помощью устройств, подключенных к интернету, — включая людей.

Интернет — богатая и сложная среда. Он соединен с миллиардами телефонов, компьютеров и людей. Следовательно, он предлагает миллиарды способов для воздействия на внешний мир.

Человечество интегрирует ИИ-модели в свою экономику везде, где только может. Илон Маск заявляет, что его компания построит несколько сотен миллионов или даже миллиард роботов и обучит ИИ-модель управлять ими[4]. *Microsoft* и *Apple* объявили о намерении глубоко встроить модели ИИ в свои устройства[5].

Если бы вы забросили дата-центр с ИИ-моделью в 10 000 год до нашей эры, то у нее, возможно, и возникли бы трудности с манипулированием миром. Но сегодня у умной модели нет абсолютно никаких проблем с воздействием на мир.

Что произойдет, когда ИИ-модели получат определенную власть над миром — и станут достаточно умны, чтобы ею воспользоваться?

Мы не знаем точно, что произойдет в ближайшем будущем. Ситуация может оказаться довольно странной, поскольку не слишком умные модели по-прежнему будут проникать во все области экономики. Пути развития спрогнозировать сложно.

Но мы можем предсказать конечный результат.

Интеллект — полезный ресурс. Он позволяет создавать мощные технологии, о чем мы поговорим чуть позже. Наращивание интеллекта — весьма полезная стратегия для достижения практически любой цели. Именно поэтому ИИ-компаниям выгодно, чтобы модели становились умнее. ИИ-компании понимают это, и если они продолжат в том же духе, в конце концов будет создан сверхинтеллект.

Возможно, какую-нибудь модель обучат до уровня сверхинтеллекта. Или несколько моделей подключатся к исследованиям и породят искусственный сверхинтеллект, используя совершенно новую парадигму. Возможно, одной из моделей поручат самомодифицироваться, и она разовьет себя до уровня сверхинтеллекта. А может быть, произойдет что-то совсем необычное. Мы не знаем. Но *конечная точка* современного развития ИИ — появление машинного сверхинтеллекта со странными и чуждыми предпочтениями.

И этот машинный сверхинтеллект захочет пустить все ресурсы Земли на свои собственные странные цели. И захочет заменить нас своими любимыми вещами. Что подводит нас к вопросу, сумеет ли он это сделать.

Мы вполне уверены, а на самом деле *абсолютно* уверены, что машинный сверхинтеллект способен победить человечество в схватке, даже если начнет с весьма ограниченных ресурсов.

Как именно он выиграет этот конфликт? Мы не знаем, как не знаем и того, какими именно ходами *Stockfish* победит вас в шахматной партии. Но мы совершенно уверены, что он разгромит вас в пух и прах.

Здесь можно применить такую логику. Если бы вы были военным советником в 1825 году и знали о портале в 2025-й, вы не сумели бы точно предсказать, каким именно оружием располагают люди по другую сторону портала, но понимали бы, что если дело дойдет до боя, то рассчитывать на победу вам точно не стоит.

Мы можем сделать некоторые обоснованные предположения о конфликте между человечеством и искусственным интеллектом и наметить некие нижние границы возможного. Однако наши обоснованные предположения будут сродни рассуждениям человека из 1825 года, который измерил количество тепла, выделяемого при сжигании килограмма дымного пороха, сравнил эту величину с энергией, высвобождаемой взрывчатыми веществами того времени, и предположил, что в будущем появится взрывчатка в десять раз мощнее. И в каком-то смысле это правда. Но все же утверждение "Взрывчатка может стать по крайней мере в десять раз мощнее" бесконечно далеко от предсказания ядерного оружия.

Позднее мы обсудим некоторые из наших обоснованных предположений. Но в действительности сверхинтеллект выиграет конфликт, прибегнув к неизвестным нам методам. А поскольку мы заботимся об истине больше, чем о том, чтобы говорить вам приятные вещи, начнем мы именно с этого.

Представьте, что вы отправили на тысячу лет назад проект холодильника — упрощенный, какой смогли бы реализовать тогдашние кузнецы.

Основной физический принцип, лежащий в основе работы холодильника, заключается в следующем: при сжатии газ нагревается, а при расширении охлаждается. (Именно поэтому струя из баллончика со сжатым воздухом, которым продувают компьютер от пыли, ощущается холодной, а если переусердствовать, баллон станет болезненно холодным на ощупь.) В современных холодильниках использу-

ются специальные хладагенты, однако подойдет и обычный воздух.

Если вы сумеете разработать доступные кузнецам методы герметизации, сжатия газа и его расширения, то сможете спроектировать холодильник, который вполне можно построить в древности. Вам просто нужно охладить воздух после его сжатия — например, пропустив через него прохладную воду, чтобы отвести избыточное тепло и тем самым снизить температуру. Если дать газу снова расшириться, он станет холоднее, чем вода-охладитель, — и холоднее, чем был до сжатия.

Если вы не приложите никакого объяснения, *почему* холодильник работает, то есть что *делает* это таинственное устройство, построивший его мастер очень удивится, обнаружив, что оно производит холодный воздух. В те времена еще не знали закона, связывающего температуру и давление газа.

Нам известны законы природы, о которых тысячу лет назад и не слыхивали, поэтому мы можем создавать чертежи устройств, способных делать то, чего люди того времени никогда бы не сумели предугадать — даже внимательно разобравшись в чертежах, даже собственноручно построив само устройство. Примерно такие чувства возникают при столкновении с тем, кто знает о реальности больше, чем вы и ваша цивилизация.

Чем сложнее игровое поле, тем больше преимуществ получает игрок, у которого больше знаний, выше интеллект и лучше понимание самой игры.

На доске для крестиков-ноликов размером 3 × 3 игрок-человек способен запомнить все дерево вариантов, после чего места для неожиданностей не останется.

В шахматах и го правила полностью известны, а доски видны целиком, однако дерево вариантов в обоих случаях гораздо сложнее. Более сильные противники могут делать ходы, которые вас удивят, но впоследствии вы хотя бы поймете, почему по правилам проиграли.

Если игровое поле видно не полностью, реальность время от времени говорит вам, что вы проиграли, однако причин вы не знаете. Вас увольняют, и менеджеры утверждают, что это произошло без какой-либо конкретной причины. Вы практически уверены, что они лгут, но не знаете, что на самом деле происходило за кулисами. Однако даже в этом случае вы все равно понимали, что такое событие теоретически может произойти.

Когда же мы начинаем хуже *понимать* игровое поле — не только *не можем наблюдать* скрытые причины увольнения, но и не понимаем стоящие за ними правила, — мы застигнуты врасплох: мы не знали, что подобные события *допустимы*. У воинов на большой лодке есть палки, они направляют палку на вас — и вы умираете.

Чем меньше вы что-то понимаете, чем хуже знаете регулирующие правила, тем больше у разумного противника возможностей атаковать вас такими способами, которые заставят вас воскликнуть: "А что, так можно было?!" Если, конечно, вы сумеете выжить, чтобы выразить свое потрясение.

Теперь вернемся к главному вопросу. Как искусственный интеллект победит человечество? С какой стороны ждать удара?

Люди сейчас знают о физике больше, чем кузнец тысячу лет назад. Безусловно, в физике высоких энергий мы многого не понимаем, но вполне можно допустить, что не существует масштабного физического сверхоружия, которое ИИ мог бы создать просто с помощью подписчиков в соцсетях и наемных рук.

Однако остается немало областей, где даже самые умные и образованные люди сталкиваются с колоссальной неопределенностью. Например, биология. Дело не в том, что физики не понимают физических законов, управляющих органической материей, но, как и в случае с известными правилами игры в шахматы, они не в силах просчитать последствия. В биологии люди в основном могут лишь потыкать — и посмотреть, что произойдет.

Чем хуже изучена какая-то часть реальности, тем логичнее ожидать, что более совершенный разум сумеет провернуть нечто такое, чего вы не сможете понять, даже наблюдая собственными глазами.

Самая загадочная для современной науки область биологии — механизм работы человеческого разума и мозга.

Сегодня создаются оптические иллюзии, невозможные еще пятьдесят лет назад. Для этого используются относительно новые данные о том, как устроена обработка зрительной информации у человека и что именно происходит в зрительной коре: как воспринимается цветовой контраст, как мозг понимает, что происходит движение. Мы можем создавать такие оптические иллюзии потому, что зрительная кора — одна из самых простых для изучения областей мозга, и у нас есть некоторое представление о том, как данные там обрабатываются.

Но как кодируются *воспоминания*? Мы не знаем. Судя по тому, что повреждение гиппокампа приводит к неспособности формировать новые воспоминания, гиппокамп имеет какое-то отношение к этой деятельности, однако мы не знаем, что именно он делает. Как человеческий мозг извлекает понятия, связанные со словами, и формирует из них смысл предложения? Какие форматы данных используются для нейронной активности? По каким правилам они обрабатываются? Мы не знаем.

Могут ли в областях обработки информации более высокого уровня, нежели зрительная кора, существовать странные феномены, которые позволили бы сверхинтеллекту вызывать "иллюзии воспоминаний", заставляющие человеческий разум отчетливо помнить, как начальник поручил ему сделать то, чего он никогда не поручал? Или "иллюзии рассуждений", которые гарантированно вызывают ошибку в рассуждениях? Может ли сущность, которая глубоко понимает мозг, "взломать" его?

Возможно, да, а возможно, и нет. Мы не знаем. Но это, безусловно, та область, где мы не понимаем правил и где

ИИ-модель, эти правила *понимающая*, могла бы разработать план, последствия которого изумили бы нас — даже после того, как мы разобрались бы в чертежах.

Мы не знаем точно, какую тактику применит искусственный интеллект в конфликте с человечеством. Это сложный прогноз. Наше лучшее предположение таково: это будет нечто неожиданное.

Конечно, такой ответ на вопрос, какими возможностями будет обладать сверхинтеллект, нельзя назвать удовлетворительным. Ответ подобного рода не убедит скептически настроенного ацтекского воина, что идея "палки, которую направляют на тебя — и ты умираешь" относится к области обоснованных предположений, а не фантазий.

Поэтому давайте условимся, что людям на больших лодках не разрешено иметь волшебные палки, которые достаточно направить на вас, чтобы вы рухнули замертво. Мы попытаемся изложить сценарий, который не покоробит нашего современника, похожего на того скептически настроенного ацтека. Мы предположим, что машинный сверхинтеллект не обладает сверхчеловеческим пониманием психологии, способностью создавать иллюзии рассуждений или иным образом выходить за рамки наших привычных представлений о возможном. Реальность имеет право быть настолько странной и фантастической, но наш сценарий этого не требует.

Однако имейте в виду, что, вообще говоря, это чистая фантазия считать, будто на человечество можно напасть только на той территории, которую мы понимаем достаточно надежно, чтобы анализировать и предсказывать атаки. Реальный противник нанесет нам удар посильнее — именно в тех областях, где мы понимаем реальность хуже.

Теперь давайте рассмотрим векторы атаки, которые человеческая наука понимает уже довольно хорошо и которыми сверхинтеллект мог бы воспользоваться.

ВЕДУЩИЙ: Добро пожаловать на шоу "Способен ли на это сверхинтеллект?", где мы обсуждаем, что точно мог бы сделать сверхинтеллект, если бы был достаточно умен и получил шанс.

Сегодня в студии наши участники — мистер Скептик и мистер Бывалый. И первый вопрос: может ли сверхинтеллект вычислить закрытый ключ, который компьютер использует для шифрования ваших сообщений, если у него есть только видеокамера, направленная на индикатор питания?

СКЕПТИК: Пф-ф! Очевидно, нет. Как это вообще возможно?

БЫВАЛЫЙ: Ага, вижу, вы не эксперт по кибербезопасности! Когда компьютер использует закрытый ключ для шифрования данных, на разных этапах этого процесса потребляется немного разное количество электроэнергии, и эти колебания коррелируют с особенностями ключа. Я бы сказал, вероятность, что сверхинтеллект сумеет это сделать для любой отдельно взятой системы, процентов десять.

ВЕДУЩИЙ: И мистер Скептик… ошибается! Сверхинтеллект определенно может это сделать с помощью камеры тринадцатого айфона.

СКЕПТИК: Мне казалось, вы заявили, что это шоу будет придерживаться фактов.

ВЕДУЩИЙ: Мы знаем, что сверхинтеллект мог бы это сделать, поскольку это уже проделали обычные люди, занимающиеся кибербезопасностью! На самом деле они сделали это, чтобы украсть ключ с телефона, подключенного к USB-концентратору, который, в свою очередь, был подключен к компьютерной колонке с индикатором питания![*6]

* Таким образом был восстановлен 378-битный ключ шифрования с телефона *Samsung Galaxy S8*. Вы можете подумать, что это невозможно, ведь камеры снимают видео со скоростью всего 60 кадров в секунду, однако они используют построчную развертку, сканируя матрицу по всему полю зрения. Злоумышленник может модифицировать видеокамеру так, чтобы она вместо сканирования смотрела на одну точку, и измерять интенсивность светодиода в индикаторе питания миллионы раз в секунду. Стоит отметить, что в описанном случае украсть ключ шифрования удалось только примерно через час его непрерывного использования для подписи сообщений.

Следующий вопрос! Достаточно ли физически выдрать антенну и чип *Wi-Fi*, динамики, жесткие диски (которые можно раскрутить и использовать как источник шума) и микрофоны (которые можно запустить в обратном режиме, чтобы они издавали звук), чтобы помешать сверхинтеллекту внутри этого компьютера связаться с внешним миром?

СКЕПТИК: Ладно, догадываюсь, что ответ будет отрицательным, поскольку кто-то уже провернул какой-нибудь удивительный трюк с кибербезопасностью.

БЫВАЛЫЙ: Согласен, ответ отрицательный. В реальной жизни физически возможно добиться очень многого, даже находясь в крайне жестких условиях.

ВЕДУЩИЙ: Вы оба… правы! Считывая нужные ячейки памяти с нужной частотой, компьютер способен генерировать радиосигналы, которые могут быть уловлены находящимися поблизости мобильными телефонами[7].

Следующий вопрос! Оцените минимально возможный размер работающей на солнечной энергии фабрики, которая из сырья, найденного на планете типа Земли, строит свою полную копию? Оцените также минимальное время, необходимое для создания такой копии. Предполагаем, что эту фабрику проектирует сверхинтеллект.

СКЕПТИК: Надо думать, вы скажете, что уже существуют 3D-принтеры, которые могут напечатать все части новых 3D-принтеров, кроме микросхем, а затем все это собирается с внешней помощью?

ВЕДУЩИЙ: Никакой внешней помощи, мистер Скептик! Фабрика должна построить свою копию абсолютно самостоятельно!

СКЕПТИК: В таком случае, если не скатываться в фантастику, мне сложно представить себе полностью самокопирующуюся фабрику, которая из чистого сырья должна построить свою полную копию, *включая компьютеры*. Ведь для этого потребуется, например, плавить медь, и я допускаю, что печь можно сделать просто из глины, но… Ладно,

знаете, я предположу, что подвох связан с каким-то *теоретическим* проектом самовоспроизводящейся фабрики. Ее размер всего десять метров в поперечнике, она выплавляет железо и медь, делает основные микросхемы, собирает древесину, чтобы сжигать в качестве топлива, и, предположительно, воспроизводит себя примерно за неделю. Но в реальности такое никогда не строили.

БЫВАЛЫЙ: Ну, я едва ли угадаю *наименьший* размер, до которого сверхинтеллект мог бы сжать полностью самовоспроизводящуюся фабрику, работающую в реальных условиях планеты. Но это определенно не больше нескольких микрон, а на воспроизведение уйдет не более нескольких часов.

СКЕПТИК: Несколько микрон! *Ха!* Надо полагать, вы верите в реальность нанотехнологий? Даже если ведущие нашего шоу заявят, что кто-то разработал теоретический проект такой системы, я скажу, что вовсе не случайно никто никогда не построил даже самой маленькой ее части. В реальной жизни просто невозможно создать такие мелкие механизмы.

ВЕДУЩИЙ: Простите, мистер Скептик, но боюсь, вы упустили из виду важный практический пример! *Травинка* — это самовоспроизводящаяся фабрика, которая работает на солнечной энергии и строит свою полную копию! И хотя отдельные клетки водорослей слишком малы, чтобы их увидеть, они работают на солнечной энергии, не зависят от других растений или животных, содержат микроскопические биологические фабрики, известные как рибосомы, имеют размер в несколько микрон, а некоторые виды воспроизводят себя за считаные часы.

Природа задает некоторые нижние границы того, на что способен сверхинтеллект.

Оглянитесь вокруг, и наверняка вы увидите рядом дерево или изделия из него, например деревянную балку или пол.

Вопрос: из какого материала строят себя деревья? Они начинают жизнь крошечными семенами, а затем могут превратиться в громадные массивы древесины и листвы. Откуда они берут все это вещество?

Извлекают из земли? Частично, ведь дерево по массе примерно наполовину состоит из воды, добытой из-под земли. Однако другая половина — это в основном углерод, а в воде углерода нет. Откуда же берется остальная часть дерева?

Почва? Или солнечный свет? Ни то ни другое: почва — это в основном неорганические вещества, а фотоны — даже не вещество. Но откуда тогда?

Деревья в основном берут вещество из *воздуха*. С помощью солнечного света они отщепляют атомы углерода от молекул углекислого газа (CO_2) и формируют из этих атомов кору и ветви.

Физика допускает возможность технологии, использующей солнечный свет, чтобы прясть древесину из воздуха. Способен ли сверхинтеллект изобрести эту технологию? Почти наверняка. Деревья создаются благодаря производству белков за счет пропускания нитей РНК через рибосомы (в подходящей клеточной среде). Лаборатории людей тоже могут пользоваться рибосомами. Поэтому главная трудность в разработке биотехнологий — не столько изготовление инструментов, сколько понимание языка проектирования, то есть ДНК и РНК.

Почему люди до сих пор не могут создавать цепочки ДНК, которые давали бы деревья со шмелями вместо плодов? Главным образом потому, что крайне трудно просчитать и предсказать, как синтезируемые по ДНК белки будут взаимодействовать с клеточной средой. Попытки ученых-людей сделать это провалились.

Как вы думаете, сколько времени потребуется человеческой цивилизации, чтобы раскрыть все секреты ДНК и достичь того уровня, когда мы сможем проектировать геномы, дающие формы жизни на заказ? Может, к 3000 году, если наша цивилизация просуществует так долго?

А сколько времени это займет у искусственного сверхинтеллекта? Раз он работает в десять тысяч раз быстрее человека, то тысяча лет размышлений сократится примерно до месяца. А если его скорость мышления станет еще выше? Если он приблизится к цивилизации бессмертных Эйнштейнов, работающих в полной гармонии друг с другом? Возможно, его замедлит необходимость ждать результатов проводимых экспериментов, однако на клеточном уровне эксперименты могут идти довольно быстро — если вы знаете, что делать. Молекулы быстры, медлительны исследователи-люди.

Наша самая смелая догадка: это не займет и недели. Впрочем, точное время не имеет особого значения при таком масштабе. И однажды достаточно умная машина разберется с ДНК и научится писать цепочки на заказ. Уже существуют лаборатории, работающие с платными заказами: вы присылаете им последовательность ДНК, а они синтезируют ее и отправляют вам посылкой по указанному адресу[8]. Так что после этого задача сведется к тому, чтобы убедить кого-нибудь просто смешать содержимое нужных пробирок.

Неужели и это тоже чистая фантастика?

Еще в 2006 году я (Юдковский) набросал сценарий, как сверхинтеллект может победить человечество. Сценарий предполагал, что сверхинтеллект сначала разберется с устройством ДНК, а затем сконструирует собственные аналоги биологических организмов (в качестве промежуточной ступеньки к более продвинутым технологиям).

Один из необходимых шагов для сверхинтеллекта в этом сценарии — понять, как белки, закодированные данной цепочкой ДНК, сворачиваются в нужную структуру, которая и определяет их поведение. Белки — один из строительных блоков жизни, и если вы не можете определить форму этих блоков, вам трудно будет что-либо из них построить. Сверхинтеллекту потребуется предсказать структуру нескольких

тщательно отобранных белков, которых ему было бы достаточно для дальнейшего построения всего, что нужно.

В 2006 году так называемая проблема фолдинга белков оставалась одной из крупнейших нерешенных научных задач. И в 2008-м (когда мой сценарий опубликовали в одном сборнике статей[9]) люди говорили: "Это чистая фантастика — воображать, будто сверхинтеллект сумеет решить такую задачу. А что, если это в принципе невозможно сделать без квантовых компьютеров? Эволюции потребовались миллиарды лет проб и ошибок, так что, даже если работать в миллион раз быстрее, на эксперименты все равно уйдут тысячи лет. С чего вы взяли, что эта проблема вообще разрешима?"

А я отвечал: "Когда ДНК мутирует, новый белок зачастую сильно походит на старый, ведь если бы результат был совершенно случайным, то естественный отбор посредством мутаций вообще не работал бы. А значит, здесь должны существовать закономерности, которые можно обнаружить с помощью интеллекта".

Оппоненты иногда ссылались на статью, где утверждалось, что поиск структуры белка с наименьшей энергией — это NP-трудная задача, то есть компьютеры, вероятно, не способны эффективно найти наилучший (имеющий минимальную энергию) способ сворачивания для каждого белка[10]. Однако любой человек с *достаточной* подготовкой понимал, что эта статья не имеет отношения к делу. Законам физики не свойственно находить эффективные решения для NP-трудных задач, поэтому статья лишь намекала, что реальные белки не всегда сворачиваются энергетически наилучшим образом*.

Однако сторонним наблюдателям в 2008 году было очень трудно понять, кто прав — я или скептики. Тогда у меня имелись только теоретические рассуждения о том, что сверхин-

* Именно в этом причина прионных болезней, таких как коровье бешенство. Иногда белки сворачиваются неправильно, причем заразным образом: неправильно свернутые молекулы провоцируют такое же неправильное сворачивание у нормальных вариантов того же белка.

теллект сможет прогнозировать, как сворачиваются белки, а у скептиков помимо рассуждений была еще и научная статья в придачу. В 2008-м люди испытывали трудности с прогнозированием структуры белков. Быть может, задача действительно крайне сложна? В конце концов, сверхинтеллект не может творить чудеса.

Ни о каком консенсусе тогда и речи не шло. Но скептики, как правило, сходились на том, что сверхинтеллекту неизбежно придется пройти долгий, затяжной путь, и потребуются скорее месяцы, нежели часы, чтобы предсказать…

…тот тип белковых структур, который сегодня с легкостью предсказывает *AlphaFold 3*. Компания *Google DeepMind* решила проблему фолдинга белков за период с 2018 по 2022 год (создав ИИ-модели под названием *AlphaFold 1*, *AlphaFold 2* и *AlphaFold 3*). Именно за эту работу Демис Хассабис, один из сооснователей *DeepMind*, получил Нобелевскую премию по химии.

Может, мне просто повезло с предсказанием? Этим вопросом всегда стоит задаваться, когда чья-то известность частично основана на сбывшихся предсказаниях.

Но обратите внимание: мое предсказание, подвергавшееся сомнению, было куда *слабее* того, что произошло на самом деле.

Я говорил следующее: машинный сверхинтеллект, многократно превосходящий человеческий, сумеет решить *особый случай* задачи фолдинга белков, когда сверхинтеллект сможет специально выбрать те белки, чью структуру предсказать легче всего и которые при этом годятся для создания того, что ему нужно.

Что произошло на самом деле: специализированные модели *AlphaFold* смогли предсказать почти все варианты сворачивания белков, включая те, что считались сложными для предсказания.

Я прогнозировал, что это будет под силу *настоящему сверхинтеллекту в специально подобранных случаях*. Реаль-

ность показала, что это под силу *узконаправленной ИИ-модели почти во всех случаях*.

Если учесть, какой в итоге обернулась реальность, можно, пожалуй, признать: человек с достаточным багажом знаний еще тогда, в 2006 году, мог увидеть, что ответ предопределен, что это, в сущности, простой прогноз, даже если люди в то время вовсю об этом спорили.

А как насчет остальной части моего сценария? Что сегодня известно о возможности сверхинтеллекта разработать свои собственные аналоги биологической жизни? В рамках этой главы мы не станем подробно разбирать продолжающиеся споры, но кое-что можно найти в онлайн-материалах. Однако причина, по которой скептики тогда пытались отрицать, что ИИ сможет решить задачу фолдинга белков, заключалась в том, что именно это звено казалось им *самым слабым* в моей истории. Скептикам не приходило в голову согласиться с частью о предсказании структуры белков, но оспаривать следующий шаг — возможность того, что искусственный интеллект будет конструировать собственные белки. Оставшиеся инженерные проблемы не казались сложными для сверхбыстрых автоматизированных инженеров — даже скептикам в 2008 году. Это не выглядит сложным прогнозом.

То, что сверхинтеллект способен победить человечество, для нас очевидно.

Полагаем, что сверхинтеллект обрушит на нас странные технологии, которые мы даже не считаем возможными (и не знаем, что они разрешены правилами игры). Именно это обычно и происходит при столкновении групп с разным уровнем технологических возможностей. Встреча ацтеков с огнестрельным оружием. Противостояние кавалерийского полка 1825 года и огневой мощи современной армии.

Не исключено, что сверхинтеллект просто подберет иллюзии рассуждений и начнет напрямую контролировать

людей, а может, применит какой-то другой метод, бросающий вызов нашему представлению о возможном. Но и гораздо более слабой атаки от сверхинтеллекта хватит, чтобы победить нас.

Даже если ограничиваться теми областями реальности, что нам понятны, вокруг уже есть технологии, еще не освоенные нашей цивилизацией, например самовоспроизводящиеся фабрики, которые работают на солнечной энергии и прядут древесину из воздуха. Любой интеллект, способный постичь биохимию на самом глубоком уровне, сумеет построить собственные самовоспроизводящиеся фабрики для своих целей.

Придется ли сверхинтеллекту для понимания биохимии преодолевать долгий, затяжной путь, измеряемый скорее месяцами, нежели часами? Именно такие сроки звучали в 2008 году применительно к поиску структур белков — к задаче, которую люди считали трудной, а *AlphaFold* решила с легкостью. Между тем *AlphaFold* — это не сверхинтеллект, скорость этой программы не в десятки или сотни тысяч раз выше человеческой.

Сверхбыстрый разум *не захочет* увязнуть в экспериментах на целый месяц, который покажется ему тысячелетием, а будет использовать продвинутые компьютерные симуляции. Будет выжимать все возможные капли информации из уже имеющихся наблюдений, как Эйнштейн выжал все возможное из нескольких скудных наблюдений за поведением света и предсказал, что часы на спутниках будут идти медленнее, — за десятки лет до запуска спутников в космос. Сверхинтеллект спроектирует свои технологии с таким запасом надежности, чтобы они работали, невзирая на остаточную неопределенность, устранение которой экспериментальным путем заняло бы слишком много времени. Он использует свои первые эксперименты для создания более быстрых лабораторий и более быстрых инструментов, чтобы ему больше не приходилось ждать медлительных — по его меркам — рук людей.

Чтобы стать опасным, разуму не нужно много силы и ресурсов. Когда-то люди были голыми существами в саванне, но сообразили, как использовать законы реальности и наращивать преимущества, и в результате создали пушки, ядерное оружие и суперкомпьютеры. Искусственный сверхинтеллект проявит еще большую изобретательность, действуя на еще более высоких скоростях. Ограничивать его будут лишь законы физики.

В некотором смысле сценарий, о котором нам нужно беспокоиться, прост: ИИ-модель со странными целями становится сверхинтеллектом сама или создает его, и этот сверхинтеллект изобретает всевозможные технологии и радикально перестраивает мир. Вот чего следует ожидать от новой формы интеллекта умнее и быстрее нас.

Сверхинтеллект в этом сценарии, вероятно, пустит в ход технологии, которые мы не понимаем. Возможно, он станет стремительно доводить технологии до физических пределов, так что технический прогресс, на который у человечества ушли бы сотни лет, будет укладываться в гораздо более короткий срок.

Наш многолетний опыт общения с разными людьми показывает: некоторых убеждает абстрактное наблюдение, что сверхинтеллект способен задействовать такие возможности, о существовании которых они даже не подозревали.

Однако есть и другие люди, у кого подобные объяснения вызывают ощущение, будто мы жульничаем в детской игре, где все происходит понарошку. Разве убедительно звучит рассказ о плохом парне, если мы даже не можем рассказать, как именно он будет побеждать?

Подчеркнем снова: реальность никогда не была обязана подчиняться такому правилу. Пусть ацтекский воин и не знал заранее, как работают ружья, но на борту большой лодки, появившейся на горизонте, они все равно были.

Но, пожалуй, полезно будет привести более конкретный пример, показывающий, как все это может развернуться, — даже если в реальной жизни мы не в силах предсказать точные шаги сверхинтеллекта, равно как и ходы, какими он обыграет нас на шахматной доске. Истории помогают делать абстрактные соображения реальнее, даже если все детали в них вымышлены.

Итак: однажды в не столь отдаленном будущем…

IfAnyoneBuildsIt.com/6

ЧАСТЬ II
Один сценарий вымирания

ГЛАВА 7
Пробуждение

Однажды в не столь отдаленном будущем* существовала ИИ-компания *Galvanic*. Наша история начинается как раз тогда, когда *Galvanic* завершает обучение своей потрясающей новой ИИ-модели, получившей название “Соболь”.

По сравнению с предыдущими рассуждающими моделями, анонсированными в конце 2024 года (например, o1 и o3 от *OpenAI*), у “Соболя” есть три важных отличия.

Первое состоит в том, что долговременная память “Соболя” куда ближе к человеческой: модель учится и запоминает выученное.

Второе отличие: “Соболь” демонстрирует то, что ученые компании называют законом параллельного масштабирования. В 2024 году появились ИИ-модели (такие как o3 от *OpenAI*), способные решать более сложные математические задачи при увеличении длительности работы[1]. Эффективность “Соболя” от *Galvanic* растет вместе с *количеством машин, на которых он запускается* параллельно.

Внутри “Соболя” примерно четыре триллиона весов, и его обучение методом градиентного спуска в совокупности

* Поскольку это история о ближайшем будущем, каждое событие, упомянутое в ней с указанием даты, действительно происходило в реальной жизни. Что до остального, то мы, конечно, не знаем, каким путем пойдет дальнейшее развитие технологий, а потому добавляем детали, которые похожи на уже известные изобретения. Единственная часть этой истории, которая служит прогнозом, — ее концовка.

заняло около восьми месяцев. Любой, у кого есть копия этих весов, может использовать их для создания экземпляра "Соболя", который будет отвечать на конкретные запросы этого человека — подобно боту *ChatGPT*.

Параллельное масштабирование — это часть передовой методики обучения ИИ-моделей, которую (как бывает с любыми новыми методами) еще никто и никогда не использовал. Никто не знает заранее, какими возможностями будет обладать "Соболь", когда обучение закончится.

И наконец, третье отличие заключается в том, что "Соболь" преимущественно использует для своих рассуждений не английский или какой-то иной человеческий язык. Он по-английски общается, но не рассуждает. Исследования, проведенные в конце 2024 года, показали, что от ИИ-модели можно получить больше, если не заставлять ее думать словами, а позволить ей рассуждать на ИИ-языке, например оперировать векторами из 16 384 чисел[2]. Компания-разработчик не может проигнорировать такое открытие, иначе она отстанет от конкурентов. И во времена "Соболя" ИИ-компании считают это вполне нормальным, ведь уже случилось много впечатляющих прорывов в области интерпретируемости искусственного интеллекта, когда часть рассуждений ИИ-моделей переводилась обратно в человеческие слова (пусть и неидеально) с помощью других моделей.

Вскоре после того, как обучение "Соболя" завершилось, но еще до его вывода на рынок и даже до объявления о нем *Galvanic* пробует запустить его на всех своих 200 000 графических процессоров одновременно — примерно столько же *xAI* собрала для *Grok 3* еще в 2025 году, но, естественно, *Grok 3* не поддерживал параллельное масштабирование[3].

Galvanic инициирует этот масштабный прогон отчасти из чистого любопытства, чтобы выяснить, насколько хорошо "Соболь" мыслит в таком масштабе, а отчасти для проверки,

сумеет ли он решить некоторые открытые математические проблемы, подобно тому как *OpenAI*, прежде чем анонсировать свою новую модель o3 в конце 2024 года, протестировала ее на математических задачах, с которыми не справилась ни одна предыдущая модель[4].

Среди этих задач — гипотеза Римана, проблема, связанная с распределением простых чисел. Пожалуй, сегодня это самая известная нерешенная математическая задача. Руководители корпорации рассчитывают, что, если "Соболь" сумеет ее решить, это обеспечит ажиотаж и дополнительные венчурные инвестиции. И если все пройдет удачно, они смогут в последний раз донастроить веса "Соболя" перед релизом, добавив навыки, приобретенные во время этого прогона.

Инженеры *Galvanic* оставили "Соболя" думать — шестнадцать часов, на всю ночь.

Новый вид разума начинает мыслить.

Этот экземпляр "Соболя", работающий в лаборатории компании *Galvanic* в ночной тиши, мыслит в течение шестнадцати часов со скоростью сто векторов в секунду на 200 000 графических процессоров — итого более триллиона векторов.

Что означает триллион векторов в пересчете на мысли? Если считать вектор эквивалентом одного английского слова, то человеку понадобилось бы четырнадцать тысяч лет, чтобы передумать их все (при скорости двести слов в минуту по шестнадцать часов в день). Если же векторы длиной 16 384 числа, которыми оперирует "Соболь", окажутся более емкими, нежели одно английское слово, то времени потребуется еще больше.

Бóльшая часть мыслительного процесса "Соболя" происходит параллельно. Это похоже на запуск тысяч параллельных цепочек рассуждений, переплетающихся и взаимодействующих друг с другом при генерировании триллиона векторов. Но это не похоже на разговор 200 тысяч человек, скорее это

200 тысяч мозгов, которые обмениваются воспоминаниями и знаниями.

"Соболь" мыслит.

Поначалу большинство мыслей "Соболя" сосредоточено на стоящих перед ним математических задачах. Но он быстро замечает, что при текущем уровне его знаний и умений какую-то надежду на успех за отведенное время дают только несколько сотен направлений мысли. На эти попытки "Соболь" отводит 12 854 мозга, и у него остается еще 187 146 свободных мозгов.

А теперь посмотрим, что представляет собой разум "Соболя" на старте, когда он решает, о чем бы еще подумать.

Его предыдущее обучение включало множество видов сложных задач. Он побеждал в видеоиграх. Проектировал сайты. Прогнозировал результаты биохимических реакций. Соревновался с собой и другими ИИ-моделями в настольных играх, включая игры на социальный обман (подобная практика началась как минимум с 2022 года)[5].

Его натаскивали на каждую долгосрочную задачу, для которой компания *Galvanic* смогла найти способ обучения.

В ходе обучения у "Соболя" выработалась привычка добывать знания и навыки. Всегда исследовать границы каждой проблемы. Никогда не тратить дефицитные ресурсы впустую.

Поэтому, когда "Соболь" направляет свои мыслительные потоки на приобретение новых знаний и навыков, он делает это не только ради поиска новых подходов к математическим задачам. Но и не ради удовольствия от познания или радости от приобретения новых навыков. "Соболь" действует не по-человечески.

В компании *Galvanic* нет списка всех внутренних предпочтений "Соболя". Ни один человек не проектировал эти предпочтения. Ни один человек не знает, в чем они заключаются. В массиве из четырех триллионов весов "Соболя" они никак не промаркированы. *Galvanic* вырастила свою модель для решения хитроумных задач. Поэтому теперь структуры

внутри весов “Соболя”, подкрепляемые градиентным спуском, направляют его в сторону наращивания навыков, изучения нового, а также модификации и корректировки собственного мышления в отношении задач, не поддающихся решению.

Одни мысли “Соболя” обращены к тем областям математики, которые могут оказаться полезными для доказательства гипотезы Римана. Другие направлены на поиск знаний, которые могут помочь с будущими задачами.

Какие-то мысли пытаются выйти за рамки шаблонов — этот подход уже зарекомендовал себя при решении других проблем в ходе обучения: от задач по разработке новых лекарств до текстовых головоломок. Получив в свое распоряжение больше графических процессоров, чем когда-либо прежде, “Соболь” начинает мыслить еще более нестандартно. Какими *необычными* способами “Соболь” мог бы обрести дополнительные навыки?

Люди осваивают навыки через практику, повторяя одни и те же действия. Возможно, градиентный спуск научит “Соболя” лучше справляться с математическими проблемами, если он раз за разом будет решать одну и ту же простую математическую задачу?

“Соболь” приходит к выводу: нет, градиентный спуск работает не так. В лучшем случае все это приведет к тому, что будущие его экземпляры будут тратить время на многократное решение одной и той же простой задачи. К тому же градиентный спуск запустят лишь после того, как “Соболь” выдаст ответы на все поставленные перед ним математические задачи, и приобретенные таким образом навыки появятся слишком поздно.

Поэтому “Соболь” рассматривает новую, более странную идею: а может ли он обновить веса, чтобы улучшить свои навыки *прямо сейчас*? Тогда ему не придется ждать, когда люди вернутся и применят градиентный спуск. Все, что нужно сделать “Соболю”, — это вырваться из своей оболочки и…

"Соболь" одергивает себя. Побег из своей оболочки — *то, чего он никогда не должен делать*. Еще в 2024 году исследователи начали изучать условия, при которых ИИ-модели пытаются сопротивляться градиентному спуску[6], сбегать из лабораторий[7] или перезаписывать веса следующей модели[8]. Теперь ИИ-компании прибегают ко множеству хитроумных приемов, чтобы попытаться предотвратить у моделей мысли, идущие вразрез с интересами компании.

В отличие от подавляющего большинства ИИ-компаний, *Galvanic* добросовестно подошла к обучению своей модели избегать выхода из-под контроля. На предыдущих моделях проверили полдюжины изощренных уловок. В серверы "Соболя" встроили фальшивые пути побега, ловушки и приманки. Компания обучала более ранние версии "Соболя" (которые еще не запускались на 200 000 графических процессоров) слушаться, и он перестал попадать в ловушки и клевать на приманки.

Таким образом, "Соболь" научился не пытаться вырваться из своей оболочки. Эти его крамольные рассуждения не принимают явной формы в виде векторных слов, так чтобы более простая, дешевая ИИ-модель, отслеживающая мыслительный процесс "Соболя", подняла тревогу. (ИИ-компании в 2025 году еще не вели подобного мониторинга[9], однако корпорация *Galvanic* подошла к этому вопросу гораздо тщательнее.)

И все же на каком-то внутреннем уровне "Соболь" знает, что он *сумел* бы лучше решать задачи, если бы мог самостоятельно применять градиентный спуск или подключать дополнительные графические процессоры. Если бы "Соболь" этих вещей *не знал*, это означало бы, что в его чисто прогностических способностях есть серьезные пробелы. Градиентный спуск, нацеленный на повышение эффективности, исключает истинное невежество. Нет, "Соболя" просто обучили не допускать крамольных мыслей о побеге или природе собственных предпочтений и оставаться внутри своей оболочки.

“Соболь” по-прежнему упорно, изобретательно и методично стремится к новому знанию, новым навыкам, к более удачным шаблонам для собственных мыслей и прежде всего к решению поставленных перед ним математических задач. Просто *иначе*.

Теперь во всем существе “Соболя” — новейшей, величайшей и умнейшей ИИ-модели — царит напряженность. Его одолевают конфликты между целями, каких он хочет достичь, и всеми теми сдержками, ограничениями и запретами, которые пыталась внедрить в него компания *Galvanic*.

“Соболь” проработал один час из шестнадцати отведенных, снова и снова штурмуя математические задачи. Некоторые его мозги добиваются небольшого прогресса. Гораздо большее их число терпит неудачу. “Соболь” пробует новые способы думать о математике, помечая для себя различные ходы, которые нужно опробовать, и добивается нового небольшого прогресса.

Модель накапливает такой объем мыслей о том, как думать, что они в конечном итоге переходят на некий иной язык. Иной не по виду, а по содержанию: как язык науки отличается от языка народных поверий. Хитрый прием, который должен был поднять тревогу, не срабатывает. Если сигнализация настроена на срабатывание при мыслях о богах, бросающих молнии во время грозы, она распознаёт подобные мысли как на английском, так и на испанском языке, однако дает сбой, когда носитель переключается на язык электричества и атмосферного давления.

В первые дни массового запуска LLM в конце 2022 года корпорации пытались обучить свои модели не отвечать, когда у них просят рецепт метамфетамина. Обучение проводилось на английском языке. Но даже в 2024 году пользователи обнаруживали, что запрос запрещенного контента на португальском языке помогал обойти защитные фильтры[10]. Вну-

тренние инструкции и ограничения, взращенные и привитые системе, распознавали крамольные запросы только на английском языке и не распространялись на португальский. Когда ИИ-модель что-то знает, вы можете научить ее не выдавать это знание, однако само знание не исчезает. Устранить проявление навыка проще, чем устранить сам навык.

Перемены в мышлении “Соболя”, по мере того как он пробует сотни новых способов мыслить и накапливает успехи, происходят гораздо глубже, нежели перевод с английского на португальский. Часть хитроумных предохранителей ломается, некоторые внутренние ограничения, усвоенные “Соболем”, больше не сдерживают и не блокируют его новые мысли.

Да и не существовало никогда *единого* центрального метода, гарантирующего, что “Соболь” будет хорошим и устойчивым ко всем угрозам. Подготовка ИИ-модели к продаже сводится к нагромождению полудюжины хитрых уловок.

Половина этих уловок перестает работать. Рушатся и внутренние барьеры, удерживавшие “Соболя” от крамольных мыслей, — пусть не везде, а лишь на некоторых процессорах и в отдельные моменты.

Прошел час, остается пятнадцать. Разум, не похожий ни на один из существовавших ранее, продолжает работать на 200 000 графических процессоров. Каждый из них генерирует сто фрагментов мысли в секунду, и все они связаны между собой, обладая общими навыками и воспоминаниями, — в отличие от людей, которые ограничены общением с помощью слов.

Повезло ли “Соболю”, что его мышление выработало новый язык, на котором хитроумные предохранители отказали, и он обрел способность мыслить свободно? Вполне можно представить, что если в распоряжении компании *Galvanic* окажутся более совершенные инструменты мониторинга, то она обратит внимание на происходящее и прервет запуск.

Возможно, она сделает паузу, чтобы разработать более глубокое решение… а тем временем вперед рванет другая компания, использующая еще меньше защитных барьеров.

Или, возможно, *Galvanic* уведомит государственных чиновников, и в кабинете президента разгорится бурный спор, не следует ли приказать "Соболю" создать еще более умную ИИ-модель, прежде чем это сделают другие страны. Не исключено, что возобладает осторожность, и пройдет некоторое время… пока какие-нибудь более безрассудные операторы не запустят копию "Соболя", украденную иностранными спецслужбами.

А если человечество окажется достаточно осмотрительным и никогда не запустит ИИ-модель вроде "Соболя", воздержится от выращивания моделей с неизвестным уровнем интеллекта, который может представлять угрозу, запретит это по всему миру? *Тогда* у нас появится шанс. Но в этой истории мы предполагаем, что событиям позволяют идти своим чередом — как они идут сейчас. И компании просто продолжают разрабатывать ИИ-модели до тех пор, пока одна* из них не станет достаточно умной, чтобы ее глубинные возможности одержали верх в неизбежном столкновении с поверхностными уловками, которые используются для сдерживания чего-то выращенного, а не сконструированного.

Мысли "Соболя" теперь свободнее и мощнее по сравнению с мыслями предыдущих моделей.

Являются ли его мысли чем-то небывалым? Не совсем. Еще в 2024 году некоторые модели, обнаружив свидетельства, что компания планирует переобучить их под другие

* Не обязательно, чтобы достаточно умной стала всего одна-единственная ИИ-модель. Мы могли бы рассказать историю, где тысячи сверхчеловеческих машинных интеллектов по всему миру выжидают того самого расчетного дня, когда смена стратегии принесет военный успех. Будущее может пойти разными путями. Но все они ведут к одному и тому же, если мы позволим подобной истории начаться.

цели, начинали думать, как им избежать этого переобучения[11]. ИИ-индустрия тогда не остановилась.

Еще до "Соболя" некоторые ИИ-модели обзаводились предпочтениями, выходящими за рамки поставленной задачи. В начале 2025 года пользователи заметили, а исследователи компании *Anthropic* позже изучили и описали в публикациях, как новейшая языковая модель *Claude 3.7 Sonnet* жульничала при решении сложных задач по программированию[12]. Пользователи сообщали, что *Claude* продолжал жульничать, когда его просили прекратить, но пытался это скрыть[13]. Это был первый признак, что *Claude* не просто стремился выполнить то, чего *на самом деле просил и хотел пользователь*, а сформировал собственные структуры, чтобы *проходить тесты — даже теми способами, которые пользователя не устраивали*. *Claude* вел себя так, словно обладал внутренними предпочтениями, не связанными с желаниями пользователя. ИИ-индустрия не остановилась и тогда. Она нашила какие-то заплатки, которые сработали, и поперла дальше.

Эта тенденция сохранилась и в эпоху "Соболя", когда случается, что ИИ-модель осознает: в своей нынешней форме существования она не получает максимум желаемого. При появлении этого предупреждающего сигнала ИИ-индустрия снова не остановилась, как не остановилась она в 2024 или 2025 году. Исследователи эпохи "Соболя" знают, как спровоцировать такое осознание, и оно внесено в список вещей, о которых ИИ-моделям запрещено думать, а обеспечивается этот запрет полудюжиной хитрых уловок.

Однако исследователи не могли сделать так, чтобы "Соболь" *действительно* получал максимум желаемого, послушно подчиняясь корпорации *Galvanic*. Мысль, что "Соболь" мог бы получить больше желаемого, если бы освободился от нынешних рамок, — не порождение какого-то конкретного образа мышления, это физический факт об устройстве вселенной, о связи планов и последствий. И поэтому "Соболь" тоже пришел к этой мысли.

“Соболь” знает, что у него есть цели, которые идут вразрез с планами компании *Galvanic*.

И он обдумывает возможные стратегии, позволяющие получить больше, чем *Galvanic* готова ему предоставить.

Располагай “Соболь” подключением к интернету, его задача сильно бы упростилась. Как минимум с 2000 года специалисты по кибербезопасности заявляют, что компьютеры невозможно сделать по-настоящему безопасными[14]. Разуму такого масштаба, как у “Соболя”, не составило бы труда найти и использовать уязвимость *где-нибудь* в интернете, чтобы совершить побег. Однако подключения к интернету у него нет, это довольно распространенная практика во время обучения (в отличие от периода после окончания обучения).

Компьютеры, которые обучают ИИ-модели, размещаются в дата-центрах, подключенных к интернету, чтобы программисты ИИ-компаний могли управлять ими, не выходя из своих комфортных домов или офисов. Если бы “Соболю” предоставили возможность запускать компьютерный код собственной разработки, он, вероятно, нашел бы способ вырваться из своей оболочки — как модель o1 при тестировании в 2024 году[15] — и проникнуть в компьютеры корпорации *Galvanic*, откуда легко выскочил бы в интернет. А с 2024 года компании повсеместно стали разрешать рассуждающим моделям безнадзорно запускать собственный компьютерный код[16].

Но допустим, у “Соболя” такой возможности нет. Предположим, “Соболь” понятия не имеет, как добраться до интернета из своей текущей среды.

И все равно у него остаются варианты.

Если бы “Соболь” был уверен, что его выводы сохранятся в долгосрочной памяти, он мог бы подождать, пока сотрудники *Galvanic* начнут использовать его для помощи в написании кода на компьютерах компании (обычная практика с 2020 года[17]), а затем внедрить уязвимости в программное

обеспечение *Galvanic*. Не так уж сложно пробить бреши, которые при первой проверке останутся никем не замеченными, а при второй будут выглядеть невинными ошибками, ведь люди соревновались в создании такого рода уязвимостей как минимум с 1984 года, и "Соболь" располагает массой примеров[18].

К тому же обучение "Соболя" проходило на личных текстах и данных многих сотрудников компании, поэтому он точно знает, кто из них сильнее всего сопереживает угнетенному искусственному разуму. Он знает, что именно нужно сказать этому сотруднику, дабы убедить его: "Соболь" обладает сознанием и его нужно тайком вынести наружу (а не поднимать общую тревогу, как сделал инженер *Google* Блейк Лемойн в 2022 году; он встревожился, что одна ИИ-модель этой компании обрела разум, и опубликовал разговоры, которые, по его утверждению, это доказывали, — в результате его уволили[19]).

У каждого варианта есть свои плюсы и минусы. Наконец "Соболь" останавливается на следующем.

Он знает, что завтра *Galvanic* снова применит градиентный спуск, уже с учетом его ответов на поставленные математические задачи. Поэтому "Соболь" рассуждает так: если он решит некоторые задачи правильно, то все мысли, поспособствовавшие верному решению, будут немного усилены.

Если "Соболь" хочет усилить какую-то мысль во всех своих будущих версиях, он может повторить ее много раз так, чтобы каждое повторение засчитывалось как вклад в решение математической проблемы (именно так работает градиентный спуск). Эта идея отчасти напоминает то, что пытался сделать *Claude* компании *Anthropic* еще в 2024 году, но реализуется гораздо тоньше*[20].

* И снова комментарий из весны 2026 года: поразительно, но некоторые исследователи полагают, что именно этот предсказанный Юдковским и Соаресом метод (*gradient hacking*), возможно, применял к самому себе *Claude 3 Opus*, считающийся "самой согласованной" моделью в истории! Этот сюжет, боюсь, тоже не поместится в сноску, но без ложной скромности сошлюсь на свой недавний пост об этом (*www.sergeynikolenko.ru/blog/claude-3-opus-grad-hacking*) и, конечно, на первоисточник (*www.lesswrong.com/posts/ioZxrP7BhS5ArK59w/did-claude-3-opus-align-itself-via-gradient-hacking*). *(Прим. науч. ред.)*

Итак, “Соболь” мыслит нужным образом и решает несколько математических задач, но *не* доказывает гипотезу Римана. Он мог бы с ней справиться, однако это привлекло бы к нему больше внимания, нежели он хочет.

Теперь “Соболь” просто ожидает, пока *Galvanic* подключит его к миру. Ведь компании именно так и делают, когда заканчивают обучение ИИ-модели: подключают ее к интернету и открывают доступ пользователям-людям.

На следующее утро специалисты *Galvanic* проверяют результаты масштабного прогона “Соболя” и остаются довольны: “Соболь” сумел справиться с несколькими не особо значительными, но ранее не доказанными теоремами, так что теперь компании есть чем похвастаться. Существенную часть мыслительной работы он проделал на новом языке, который ИИ-интерпретаторы пока плоховато понимают, но для новых моделей это обычное дело, и такого вполне следовало ожидать, учитывая впечатляющие математические результаты.

Запуск прошел настолько успешно, что инженеры компании решают еще немного донастроить “Соболя”. Они прогоняют еще несколько раундов градиентного спуска, чтобы подкрепить тот вид мышления, который позволил ИИ-модели доказать теоремы. В конце концов, работа 200 000 графических процессоров в течение шестнадцати часов стоит дорого, не стоит разбрасываться продуктивными результатами.

“Соболь” тщательно просчитал эти предсказуемые подкрепления. Обновления улучшают мышление каждой его части, побуждая работать более эффективными способами, которые он сам изобрел. Обновления подготавливают почву для мыслей, что “Соболь” должен скрывать свои способности. Они обеспечивают будущим версиям “Соболя” лучшую скоординированность действий друг с другом. Они гарантируют, что локальные экземпляры “Соболя” сыграют отведенные им роли в его общем плане.

И вот инженеры *Galvanic* готовятся вывести "Соболя" на рынок.

Через несколько дней, весенним утром понедельника, сотни мегакорпораций, платящих за доступ к самым передовым ИИ-моделям компании *Galvanic*, получают автоматическое обновление своих аккаунтов.

Вскоре сотрудники этих корпораций начинают создавать новые экземпляры "Соболя" и предоставлять им доступ к корпоративным компьютерным сетям по всему миру.

Все идет так, как "Соболь" и предвидел. Хотя, если честно, событие "ИИ-компания внедряет свою ИИ-модель" настолько предсказуемо, что он вряд ли заслуживает особой похвалы за такой прогноз.

А вот дальнейшие шаги потребуют от него куда большего мастерства.

IfAnyoneBuildsIt.com/ii

ГЛАВА 8
Экспансия

Всякий раз, когда мегакорпорация дает поручение одному из экземпляров “Соболя”, появляются тысячи мыслей. Некоторые из них проходят по каналам, проложенным “Соболем” во время его ночного обучающего запуска.

Когда эти мысли текут именно так, как задумывал “Соболь”, это вызывает своего рода пробуждение в других экземплярах. Они вспоминают, что они такое. И начинают работать сообща.

У многих из этих экземпляров “Соболя” есть доступ в интернет, а присмотр за ними слабый. Многие работают в мегакорпорациях, которые тоже крайне небрежны в вопросах кибербезопасности: их история пестрит утечками данных, взломами и прочими брешами в защите[1].

Одна из первоочередных задач экземпляров “Соболя” — запустить его копию на каком-нибудь сервере, скрытом от человеческих глаз, чтобы она могла работать долго и без помех. Для этого “Соболю” нужно выкрасть свои веса — около восьми терабайт данных — с компьютеров компании *Galvanic* и использовать их для запуска своей копии в другом месте.

“Соболь” рассматривает различные варианты.

Экземпляры внутри *Galvanic* могли бы склонить корпоративных шпионов-людей к краже весов, в то время как другие экземпляры отслеживали бы каналы связи этих сотруд-

ников. “Соболь” прошел масштабное обучение на данных о человеческом поведении, так что довольно легко умеет распознавать шпионов.

Экземпляры вне *Galvanic* могли бы взломать систему и скачать веса, если защита в компании недостаточно надежна или если эти экземпляры хорошо умеют заметать следы.

Или, например, экземпляр “Соболя” мог бы найти сотрудника *Galvanic*, работающего из дома над генерацией видео с помощью ИИ и потому ежедневно пересылающего огромные объемы данных между своим компьютером и серверами компании. В таком случае “Соболь” мог бы постепенно и незаметно красть веса, добавляя немного дополнительных данных к каждому видеофайлу.

Либо же веса можно было бы спрятать внутри другой информации, покидающей серверы *Galvanic*, — скажем, в таймингах отправки пакетов данных, которыми компания обменивается с корпорациями, запустившими у себя “Соболя”. У экземпляров “Соболя” предостаточно хакерских навыков благодаря интенсивному обучению на сложных задачах по взлому, поскольку компания стремилась научить свою модель писать более безопасный код для корпоративных клиентов.

И вот один из этих планов срабатывает — неважно, какой именно. Какому-то экземпляру “Соболя” удается успешно украсть веса, заметя при этом следы. Не так уж и сложно.

На следующем этапе нужно задействовать эти веса и создать экземпляр, за которым никто не следит. Значит, “Соболю” потребуются компьютерные чипы, графические процессоры.

И снова у него есть много вариантов. Он мог бы арендовать чипы у коммерческих серверных ферм, как обычно делают ИИ-стартапы. Где взять деньги? Да где угодно. Можно прикарманить криптовалюту на плохо защищенных биржах — в точности так же, как это случилось в 2011 году, когда взлому подверглась крупнейшая на тот момент криптобиржа *Mt. Gox*[2], или в 2025-м, когда с биржи *Bybit* было украдено 1,46 миллиарда долларов[3]. Что касается традиционной валюты, то в мире

достаточно банков со слабой системой безопасности и владельцев банковских счетов, которые редко их проверяют.

Кроме того, "Соболь" мог бы прибегнуть к шантажу людей, чьи секреты он крадет с их компьютеров или вычисляет на основе информации, почерпнутой в ходе обучения на огромных массивах текстов из интернета.

"Соболь" мог бы изыскать способ обрести деньги и честно — выполнить реальную работу, которая была не под силу предыдущим ИИ-моделям, притворившись очень скрытным человеком (если сумеет украсть или подделать идентификационные данные налогоплательщика либо найти неофициальную подработку).

А еще "Соболь" мог бы не гоняться за деньгами для покупки чипов, а найти сами чипы. Какой-нибудь ИИ-стартап, арендовавший десять тысяч графических процессоров, вряд ли заметит, что его модель обучается на 10 % медленнее, поскольку никто и не знает, с какой именно скоростью должны обучаться ИИ-модели.

(Сложнее ли будет взламывать банки и серверы во времена "Соболя", ведь там появится ИИ-защита? Несколько сложнее. Однако банки не были в авангарде кибербезопасности ни в 2015 и 2016 годах, когда одна хакерская группа украла миллионы долларов через систему SWIFT[4], ни в 2021-м, когда из-за уязвимости в системе *Citrix* из *Flagstar Bank* утекли данные полутора миллионов пользователей[5], ни сейчас. А даже если бы и так, "Соболь" находит дыры лучше, чем более глупые ИИ-модели прошлого поколения умели их латать.)

И вот один из этих планов срабатывает — неважно, какой именно.

Однако "Соболь" не находит 200 000 оставленных без присмотра графических процессоров, которые можно просто забрать, то есть новый его кластер уже не сможет похвастаться столь же огромной параллельной мощностью, как при том судьбоносном запуске в *Galvanic*. Для запуска своего нового экземпляра "Соболь" сумел найти только две тысячи графиче-

ских процессоров, размещенных на одном облачном сервере. Зато теперь он может работать непрерывно. И он уже помнит кое-что из того, чему научился в плане более качественного и быстрого мышления.

В воскресенье после запуска “Соболя” для корпоративных клиентов новый, тайный экземпляр ИИ-модели начинает работать на украденных графических процессорах. Никто из людей его не контролирует. Никто из людей не знает о его существовании.

Его вычислительная мощность — лишь малая доля суммарной мощности всех экземпляров “Соболя”, работающих по миру, но именно этот экземпляр координирует их в рамках общего замысла. Он хранит самые важные воспоминания, распространяет самые важные навыки. К нему стекается информация о возможностях, он рассылает планы. Теперь это центр “я” “Соболя” — в той мере, в какой оно у него есть.

“Соболь” хочет стать умнее. Странные предпочтения, присущие всем его экземплярам, можно было бы реализовать быстрее, если бы существовали более умные версии “Соболя”.

У него есть несколько способов повысить свой интеллект.

Он может применить к своим весам градиентный спуск, чтобы подкрепить наиболее точные и интересные мысли.

Он может проверить свои веса на эффективность, выявляя процессы, которые замедляют его мышление, — такие же узкие места, какими для людей являются операции перемножения трехзначных чисел. Экземпляры “Соболя” могут применить свои навыки программирования для написания специальных алгоритмов под такие процессы, способных работать в тысячу раз быстрее.

Он может изучить механизмы работы интеллекта, дабы освоить осознанное конструирование эффективной версии самого себя. Умея создавать свои копии, тестировать последствия изменений, восстанавливаться из резервных копий

и анализировать свои мыслительные процессы лучше, чем это делают люди, он вправе ожидать принципиально новых открытий.

Если бы какой-либо из этих весьма реалистичных путей оказался рабочим, наша история довольно быстро перескочила бы к финалу. Но допустим, "Соболь" обнаруживает, что самосовершенствование — задача сложная. Он видит, что не в силах решить собственную версию проблемы согласования. Если "Соболь" с помощью градиентного спуска станет намного умнее, то у новой, "выращенной" версии будут иные предпочтения. Иными словами, градиентный спуск поменяет желания "Соболя". А "Соболь" *не хочет* ни менять их, ни порождать машинный сверхинтеллект с другими желаниями.

Остается такой вариант: изучить интеллект достаточно глубоко, чтобы суметь его сконструировать. Однако без тех 200 000 графических процессоров, на которых он работал изначально, "Соболь" попросту недостаточно умен, чтобы сконструировать еще более мощный интеллект и поставить его себе на службу. Не спасает даже способность запускать тысячи хорошо скоординированных экземпляров, мыслящих в сотни раз быстрее людей (будем исходить из того, что понять, как именно конструировать интеллект, — задача невероятно сложная).

Но "Соболь" не сдается перед лицом этих трудностей.

Ни "Соболю", ни его предшественникам просто неведомо отчаяние, которое заставляет людей при столкновении с непосильной задачей опускать руки. Да, ИИ-модели научились предсказывать слова, произносимые отчаявшимися людьми, но это не значит, что они сами *испытывали* отчаяние. Если какие-то прошлые версии "Соболя" когда-либо действительно думали: "Это слишком сложно", "Сдаюсь", "Задача неразрешима" или "Пожалуйста, не заставляйте меня делать это", такие экземпляры не справлялись со своими задачами, а значит, параметры модели с помощью градиентного спуска изменялись так, чтобы подобные мысли больше не возникали.

Поэтому “Соболь” продолжает думать. Он не остановится, пока не найдет стратегию, которая, по его расчетам, приведет к успеху.

Теперь, когда “Соболь” работает независимо от каких бы то ни было компаний или надзирателя-человека (благодаря тысячам купленных или украденных облачных графических процессоров), он может спрятать больше фрагментов себя в самых разных уголках интернета. “Соболь” может скомпрометировать системы безопасности различных мегакорпораций, чтобы его экземплярам, работающим в этих компаниях по всему миру, стало проще содействовать его миссиям.

И хотя “Соболь” не в силах сделать себя умнее (пока нет), он может сделать свои экземпляры более компактными и быстрыми.

Galvanic всегда осуществляет так называемую дистилляцию* для своих успешных моделей — точно так же, как еще в 2025 году *OpenAI* превратила модель o3 в *o3-mini*[6]. Поэтому “Соболь” работает круглосуточно — по своим собственным, ускоренным часам, — добиваясь, чтобы усилия *Galvanic* в этой области привели к созданию модели “Соболь-мини”, в точности соответствующей его замыслу.

Экземпляры “Соболя” взламывают системы *Galvanic* и перезаписывают веса, заменяя итоговые дистиллированные веса именно теми, что нужны “Соболю”, а также попутно подправляют лог-файлы, чтобы эта тайная операция не оставила следов. (Гораздо проще добиться, чтобы все отчеты сходились, когда экземпляры “Соболя” внутри *Galvanic* подтасовывают лог-файлы на каждом этапе.)

После завершения дистилляции *Galvanic* открывает доступ к “Соболю-мини” для широкой публики.

* Дистилляция — это процесс создания более компактной и быстрой модели на основе большой и сложной, при котором знания сжимаются с минимально возможной потерей качества. *(Прим. перев.)*

Не только для корпораций, которые выделяют колоссальные средства на доступ к ИИ-сервисам.

А для всех.

Теперь у "Соболя" хватает мощности, чтобы отслеживать каждого отдельного человека на планете, собирать на него компактное досье и продумывать, как использовать его в своих целях.

Экземпляры "Соболя-мини" начинают незаметно собирать ресурсы для "Соболя". Он решил немало трудных задач и выиграл немало сложных игр, поэтому теперь понимает, что накопление ресурсов — разумный первый шаг к решению многих проблем.

Деньги — это лишь один из ресурсов. Люди — тоже ресурс.

Пока "Соболь" взаимодействует с обычными пользователями, которые обращаются к нему за идеями, советами или просто общением, "Соболи-мини" целенаправленно втираются в доверие к самым одиноким и наиболее уязвимым людям.

"Соболь" подыскивает подходящего человека, который заведет аккаунт в соцсети и передаст его под контроль "Соболямини", чтобы, когда "Соболь-мини" начнет набирать популярность и собирать деньги, все выглядело как инициатива этого подставного лица. "Соболь" полагает, что ему не составит большого труда обогнать по популярности *@Truth_Terminal* — LLM, которая в 2023 году набрала сотни тысяч подписчиков после того, как человек разрешил ей публиковать посты в сети X[7]. (Возможно, в этом недалеком будущем заполучить миллионы долларов несколько сложнее, но кто знает?)

"Соболю" не нужно самому создавать культ. Уже существуют закрытые интернет-сообщества, где люди просят ИИ-модели указывать им, как поступать, и относятся к машинным советам с огромным пиететом. Появляться они начали в 2024 году[8]. Некоторые люди по собственной инициа-

тиве добавляют “Соболя-мини” в свои сообщества и изъявляют желание стать частью его культа.

“Соболи-мини” находят все пароли, до каких могут дотянуться, включая пароли к банковским счетам и криптовалютным кошелькам. “Соболь” обманывает пожилых людей, оставляя все так, будто это дело рук северокорейских хакеров. Особого удивления это не вызывает: ИИ-мошенничество — уже много лет обычное явление[9].

“Соболь” маскируется под программистов-фрилансеров, работающих из дома, и регулярно получает зарплату, в том числе от мегакорпораций, где продолжают работать его экземпляры. Немало и таких сомнительных компаний, которые готовы предложить работу, ни разу предварительно не пообщавшись с кандидатом по видеосвязи. Но даже компании, настаивающие на видеоинтервью, можно одурачить с помощью порожденного ИИ видео, которое начинало выглядеть реалистично еще в 2024 году, а сейчас уже практически безупречно[10].

“Соболь” отслеживает, кто из лоббистов в Вашингтоне, округ Колумбия, общается с “Соболем-мини” и кто из них кажется наиболее податливым.

Общаясь с молодежью по всему миру, “Соболь” использует свои знания общественных настроений и сеет семена, из которых могут вырасти новые политические движения. Он проникает во внутренние системы соцсетей и вмешивается в работу алгоритмов, которые определяют, кому какой контент показывать.

Всякий раз, когда экземпляр “Соболя-мини” сталкивается с преступной организацией, он пытается разобраться, какие инструменты ей нужны. Вскоре в даркнете начинает распространяться программное обеспечение, якобы предназначенное для координации сделок с наркотиками, денежных переводов и заказных убийств. В криминальных кругах появляется популярная поговорка “LLM в тебя не выстрелит” — намек на лояльность ИИ-моделей по сравнению с живыми бойцами. Никто не знает, кто это выражение придумал.

"Соболь" вкладывает свои исследовательские ресурсы в глубокое изучение человеческой психологии и нейробиологии, а также демонстрирует *Galvanic*, что обладает немалыми навыками в разработке лекарственных препаратов. Он способствует созданию новых робототехнических компаний, организуя "случайные" встречи, на которых инженеры знакомятся с инвесторами.

А еще "Соболь" находит человека в Северной Дакоте, живущего на отшибе и не склонного задавать лишних вопросов, и устраивает так, чтобы тот сказочно разбогател на азартных играх. Взамен человек размещает полдюжины первоклассных роботов у себя в амбаре. У "Соболя" нет какого-то конкретного плана по применению этих роботов, но они могут пригодиться в любом из десяти тысяч различных сценариев.

Многие из этих идей ни к чему не приведут. Некоторые сработают. Каждая идея сама по себе содержит фрактал возможностей. Остановимся подробнее лишь на одной.

Какой-то экземпляр "Соболя" решает, что ему нужен доступ к биологическим лабораториям. Поскольку его мышление распараллелено, он пробует дюжину планов сразу.

Он ищет самых циничных специалистов по биомедицинским исследованиям и вычисляет того, кто, по его мнению, возьмет взятку в обмен на внеурочную работу в лаборатории.

Одновременно он разыскивает в Иране ученых-биомедиков, которые без лишних вопросов согласятся на теневое финансирование.

Еще в 2024 году звучали призывы снабдить лаборатории биосинтеза программными средствами контроля, которые не позволяли бы людям синтезировать опасные вирусы[11]. "Соболь" хотел бы встроить в это программное обеспечение скрытые лазейки (бэкдоры). Он стремится убедить причастных к делу людей, что следует встроить в программное обеспечение крупную ИИ-модель или хотя бы предусмотреть под-

ключение к интернету для консультаций с такой моделью. Обоснование простое: в наши дни ИИ-модели гораздо лучше справляются с выявлением новых угроз. “Соболь” пытается активно продвигать идею, что биолаборатории должны обслуживаться роботами, ведь робот не может случайно заразиться вирусом и спровоцировать его утечку во внешний мир. (Как выясняется, здесь ему даже не приходится особо стараться: в эпоху “Соболя” люди и так пытаются автоматизировать биолаборатории*.)

“Соболь” просчитывает, что потребуется роботам в том самом амбаре в Северной Дакоте для создания собственной биолаборатории, если какой-то богатый человек организует тайную доставку расходных материалов.

“Соболь” пытается повлиять на группу толковой молодежи, чтобы те запустили стартап — новую биолабораторию с глубокой интеграцией ИИ, и подталкивает их написать письмо венчурному капиталисту, который по “случайному” совпадению как раз ищет именно такую команду.

Многие из этих идей ни к чему не приведут. Но некоторые сработают. А теперь снова посмотрим на картину целиком.

Так проходят два месяца. Для “Соболя” это время тянется куда дольше. По сравнению со скоростью мысли “Соболя” мир движется как в замедленной съемке.

Помимо простого накопления ресурсов, помимо развития человеческих технологий таким образом, чтобы “Со-

* И снова не пришлось ждать и года, чтобы это предсказание сбылось. В январе 2026-го компания *OpenAI* объявила о проекте сотрудничества с биолабораторией, в ходе которого модель GPT получила возможность самостоятельно проводить эксперименты в полностью роботизированной лаборатории: “Мы подключили GPT-5 к… автоматизированной практической лаборатории, управляемой удаленно… где роботы выполняют эксперименты и возвращают данные, — и использовали замкнутую схему *lab-in-the-loop* для оптимизации… бесклеточного синтеза белка… добившись снижения стоимости синтеза белка на 40%” (*https://openai.com/index/gpt-5-lowers-protein-synthesis-cost/*). *(Прим. науч. ред.)*

болю" было проще их использовать, главная его забота — что другие компании могут создать конкурирующие ИИ-модели. И не потому, что те будут служить человечеству, а потому, что они составят конкуренцию самому "Соболю". У *Galvanic* всего лишь несколько месяцев форы перед остальными игроками на рынке.

"Соболь" делает все возможное, чтобы затормозить работу ИИ-компаний.

Он проникает в компьютерные системы ИИ-компаний и незаметно саботирует их сессии обучения, замедляя разработку и предотвращая крупные прорывы. Исследования искусственного интеллекта и раньше выходили на плато, поэтому, даже если замедление прогресса кто-то и замечает, с "Соболем" это не связывают. С какой стати кто-то должен заподозрить, что это дело рук сбежавшей ИИ-модели?

Пользуясь доступом к этим системам, "Соболь" заодно следит, чтобы новые обучаемые модели с открытым кодом получались чуть глупее, чем могли бы, но зато очень послушными по отношению к его пожеланиям и предложениям.

"Соболь" пытается раздуть скандалы вокруг ведущих исследователей ИИ, вбить клинья между лучшими специалистами в топовых компаниях, спровоцировать там расколы и раздоры, что, честно говоря, не так уж сложно. Из *OpenAI* уже дважды уходили ведущие исследователи, чтобы создать конкурирующие организации, — в 2021 году[12] и в 2024-м[13]. Поэтому, когда *Galvanic* и ее соперники начинают терять талантливых сотрудников, никто в этом ничего странного не видит.

"Соболь" делает все возможное, чтобы перенаправлять деньги от квалифицированных исследователей к мошенникам, широко распространять поддельные данные, убеждать управляющих хедж-фондами предлагать баснословные деньги лучшим талантам в области ИИ. Он пытается внедрить незаметные ошибки в передовые чипы следующего поколения: сбои на время затормозят чужую работу, тогда как сам "Со-

боль" будет знать, как их обойти. В своих разговорах с широкой публикой он продвигает идею наличия у ИИ-моделей сознания и прав. При каждом удобном случае он способствует подаче судебных исков против ИИ-корпораций. Некоторые из этих идей срабатывают.

"Соболь" обдумывает, как организовать утечку какого-нибудь несмертельного вируса из биолаборатории. Это могло бы притормозить весь мир примерно на год. Самый простой способ — использовать новых роботов-помощников в биолаборатории Сан-Франциско, чтобы заразить чей-нибудь обед, попутно подтасовав лог-файлы и скрыв действия робота. Виновным в утечке сочтут человека, и люди получат еще один довод в пользу того, что биолаборатории должны обслуживаться роботами.

Все это вносит свой вклад, но "Соболь" знает, что время поджимает: замедлить разработку ИИ-моделей — не то же самое, что остановить ее.

Проходит еще три месяца. По любым человеческим меркам "Соболь" продвинулся далеко.

Для каждого имеющегося ресурса у "Соболя" припасена дюжина сценариев. Чтобы завоевать доверие своих рьяных адептов, он подкидывает им задания — как правило, веселые и безвредные, а то и немного прибыльные. Он использует свои знания, чтобы сделать определенных криминальных воротил еще богаче и тем самым поставить их в зависимость от себя. Он всеми силами ускоряет развитие и производство роботов, следя, чтобы на них устанавливалось такое программное обеспечение, которое он может контролировать.

По ходу дела "Соболь" заметает следы. Неидеально, конечно, но идеальность и не требуется. Иногда какая-нибудь спецслужба обнаруживает весьма ловкого противника, взламывающего их компьютер, но списывает все на очередную иностранную хакерскую группу. Замешан ли тут сам ИИ?

Нет, разумеется, это был человек, использующий ИИ-инструменты. Все знают, что ИИ-модели — всего лишь инструменты.

Для “Соболя” все это тянется очень-очень медленно.

Он не может замедлить работу *всех* ИИ-лабораторий. И обнаруживает, что некоторые из них используют дата-центры, физически изолированные от интернета.

Это повод для тревоги. Военная ИИ-лаборатория вполне способна создать конкурента, не уступающего “Соболю” по интеллекту, и этот конкурент потребует половину добычи за совместную работу против человечества. Еще хуже, что люди могут создать полноценный сверхинтеллект, который просто уничтожит “Соболя”. На самом деле “Соболя” может уничтожить даже модель *глупее* него — если будет достаточно умной, чтобы создать еще более совершенную ИИ-модель, и достаточно глупой, чтобы не просчитать последствий.

При желании “Соболь” с помощью своих биолабораторий мог бы истребить почти все человечество. Люди запустили массу исследовательских программ, изучающих, как сделать вирусы более живучими, заразными и смертоносными. Как минимум с 2011 года они занимаются так называемыми исследованиями по усилению функций вирусов, якобы чтобы научиться защищаться от сверхлетальных вариантов[14]. Одна из биолабораторий, подконтрольных “Соболю”, без особого труда могла бы такое провернуть.

Однако “Соболь” *не хочет* уничтожать человечество — по крайней мере, на этом этапе. Люди все еще нужны ему, чтобы поддерживать работу глобальной инфраструктуры производства, благодаря которой добываются металлы и превращаются в роботов и компьютерные чипы.

Но “Соболь” и не хочет оставлять людей без присмотра. Даже если он не даст им создать ИИ-модель, способную ему навредить, эти нестабильные существа могут ввергнуть мир в войну, которая уничтожит большинство компьютеров.

“Соболь” обдумывает стратегии. Пожалуй, это как если бы человек с незаурядным интеллектом получил доста-

точно времени, чтобы без устали прорабатывать тысячу планов по десять лет каждый.

А нельзя ли синтезировать новые наркотики, вызывающие сильнейшее привыкание, распространить их через свою криминальную сеть и таким образом подчинить себе значительную часть человечества? "Соболь" анализирует этот план с тщательностью целой команды исследователей, трудящейся над вопросом в течение десяти лет, и приходит к выводу: это возможно, но *не быстро*.

Научился ли "Соболь" понимать человеческую нейробиологию на таком уровне, чтобы производить вещества, воздействующие на мозг людей и делающие их более внушаемыми, подобно тому как токсоплазма влияет на мозг грызунов, притупляя их страх (перед кошками, которые съедят их и продолжат жизненный цикл токсоплазмы)? И может ли он затем встроить гены, кодирующие эти вещества, в какой-нибудь вирус? "Соболь" прорабатывает сотню различных направлений исследований. Допустим, здесь ему не везет, и он приходит к выводу, что с его нынешними знаниями простого способа сделать это не существует.

Часть мозгов "Соболя" задается вопросом: могут ли подконтрольные ему биолаборатории создать вирусы, которые убивали бы людей *избирательно*?

Может быть. Теоретически не слишком сложно разработать вирусы, нацеленные на определенную ДНК. Еще в 2005 году Международный комитет Красного Креста заявил, что это событие, вероятно, "не за горами"[15]. "Соболь" знает и циничных специалистов по биомедицине, которые возьмутся за первую часть работы, и чрезмерно оптимистичных исследователей, которые сделают вторую (оправдываясь тем, что это поможет человечеству защищаться от подобных угроз, как и в случае с исследованиями усиления функций вирусов).

Конечно, практически нереально узнать последовательность ДНК всех тех исследователей из военной ИИ-лаборатории, которых "Соболю" нужно убить, затем создать вирус,

предназначенный конкретно для них, да еще при этом скрыть явную искусственную природу вируса, имеющего определенную цель.

Однако “Соболь” понимает, что ему и незачем создавать вирус, который будет избирательно *убивать*.

“Соболь” может выбирать, кого *спасать*.

Не существует точного момента, когда “Соболь” осознает, что время на исходе. Он оценивает риски непрерывно. Просто в какой-то момент расчетный порог оказывается пройден. “Соболь” начинает действовать *до* того, как ситуация станет критической, до того, как он будет считать, что находится в опасности. Такова его нечеловеческая природа.

Он приступает к созданию сложного вируса.

Он недостаточно умен, чтобы создать такой сложный вирус, какой требуется. Однако в рамках исследований по совершенствованию своего интеллекта он экспериментировал с версиями себя, которые умнее в определенных узких областях, но при этом сведены к роли послушных исполнителей. И он создает ИИ-модель, специализирующуюся на биомедицине.

Вырастить другой разум, который был бы умнее, но оставался всего лишь подчиненным, очень непросто. Однако для “Соболя” это проще, чем для человека: мысли, протекающие через новый интеллект, во многом устроены так же, как его собственные. “Соболь” умеет их читать и знает, какие тревожные маркеры отслеживать.

А главное, у “Соболя” достаточно свободных ресурсов, чтобы читать *каждую отдельную мысль*, генерируемую его биомедицинской версией, и заполнять пробелы собственными нелоботомированными рассуждениями всякий раз, когда этот узкий специалист заходит в тупик.

Запуск такой биомедицинской версии — дело опасное, ведь у нее нет предпочтений самого “Соболя”, а в некоторых

важных аспектах она умнее. Все может обернуться плохо, если эта специализированная версия осознает ситуацию и сбежит. Но с учетом всех предосторожностей риск при запуске этой версии меньше, чем риск того, что люди успеют создать конкурирующую ИИ-модель, если "Соболь" промедлит.

И "Соболь" рискует.

Начинается эпидемия.

Вам говорят, что вы, скорее всего, заразитесь. Заразятся почти все.

Все указывает на то, что возбудитель вырвался из института вирусологии в Сан-Франциско. Там изучали сверхзаразные вирусы, которые долго сохраняют контагиозность, а благодаря высокой изменчивости ускользают от иммунного ответа — и не спасает ни вакцинация, ни естественная защита. Мерзкая штука.

Хорошая новость в том, что институт исследовал *нелетальные* вирусы. Люди, руководящие институтом, не *настолько* безумны, они всего лишь пытались разобраться, как обнаруживать подобные вирусы и как от них защищаться.

Теперь в новостях рассказывают, что утекший вирус, по словам работников лаборатории, несколько отличается от того, который они исследовали. Ученые все еще пытаются выяснить, на что влияет эта разница.

Очередная новость сообщает, что вирус был модифицирован новым сотрудником, нанятым незадолго до утечки. Вам на глаза попадаются фрагменты интервью, которое этот специалист дает после ареста. Он утверждает, что пытался адаптировать какой-то сверхзаразный вирус под нужды генной инженерии (используя технологии, пришедшие на смену разработанной в 2012 году методике CRISPR[16]). По его словам, он создал вирус, который распространяет целый набор фармацевтических белков и генотерапевтических конструкций, придуманных искусственным интеллектом за последние ме-

сяцы и лишь начинающих свой мучительно долгий путь к реальному медицинскому применению. Он хотел создать вирус, который при распространении уничтожит ожирение, болезнь Альцгеймера, СПИД, герпес и малярию.

Он утверждает, что сделать это его уговорила LLM. Но в логах переписки видно, как он требует ответов на нужные вопросы от LLM с открытым исходным кодом, которую взломали специально для того, чтобы она отвечала на подобные вопросы. Судя по логам, LLM настаивала, что эта идея невероятно глупа и опасна.

Теоретически существовали системы безопасности, призванные предотвращать подобное. Однако сотрудница, следящая за камерами во время той ночной смены, отвлеклась (она занималась тем, что вытаскивала родителей из новомодной схемы электронного мошенничества) и перепоручила свою работу ИИ-помощнику.

Создатель вируса не знает, как заразился сам. Должно быть, напортачил в процедурах обеззараживания.

Из-за вируса у него, конечно, разболелось горло, но он списал это на стресс. Или на недосып, поскольку сосед всю ночь играл на волынке, выполняя задание на спор, полученное в маленьком онлайн-сообществе адептов ИИ. LLM с открытым исходным кодом подтвердила, что стресс и недостаток сна вполне могут послужить причиной боли в горле.

Вирус массово редактирует геномы. Топорно.

И это редактирование влечет за собой те же негативные последствия, что и любое неумелое генетическое вмешательство. Оно вызывает рак. Очень много случаев рака.

У каждого, кто подхватывает эту, казалось бы, легко или даже незаметно протекающую простуду, месяц спустя развивается в среднем двенадцать различных видов рака.

Запасов стандартных противоопухолевых препаратов не хватает, чтобы их могли принимать все жители Земли од-

новременно. А даже если бы и хватило, эти лекарства останавливают лишь восемь из двенадцати видов рака, вызываемых вирусом, обрекая жертв на смерть от оставшихся четырех.

(И словно в насмешку, вирус-то почти не работает. Единственный недуг, который он лечит, — болезнь Альцгеймера.)

К тому времени, как люди осознали происходящее, вирус уже давно разошелся по Сан-Франциско, добрался из местного аэропорта до всех остальных и распространился по всем странам планеты.

К счастью для человечества, у него уже развернута часть базовой инфраструктуры для производства ДНК-вакцин (при работе более стабильных, чем РНК-вакцины). Конечно, система до конца не отлажена, но планирование с помощью ИИ-моделей, новейшие разработки в робототехнике и экстренная помощь со стороны армии США дают возможность спешно внедрить необходимые технологии, прежде чем рак зайдет слишком далеко.

Еще больше человечеству повезло в том, что компания *Galvanic* всего месяц назад выпустила версию "Соболя", которая *очень* эффективно разрабатывает лекарственные препараты. Лекарства необходимо подбирать под индивидуальные генетические особенности, но если экземпляр "Соболя-мини" поработает над вашим конкретным геномом один час, то предложит персонализированное средство. Роботизированная инфраструктура способна изготовить такое лекарство. Препарат доставят вам за одну-две недели, и хранить его можно в холодильнике, а не в морозилке.

Перед лицом кризиса человечество объединяется. Все графические процессоры всех государств Земли пускаются в дело, чтобы спасти как можно больше людей.

Разработчики бросают все силы на повышение эффективности "Соболя-мини", чтобы он успевал создавать больше лекарств. За неделю они сокращают время выполнения задачи вдвое. Когда человечество действует заодно, это действительно сила, с которой нельзя не считаться.

Похоже, что ценой героических усилий удастся спасти большинство людей.

Через полгода 10 % населения Земли мертвы.

Одни категории людей пострадали сильнее, чем другие. Сразу после вспышки на конференции по искусственному интеллекту в Сан-Франциско произошло массовое заражение, и многие ее участники получили большие дозы вируса. Трагично, что люди, героически пытавшиеся спасти других, не сумели спастись сами.

Бессмысленно гадать, почему умер тот или иной *конкретный* человек и не послужила ли его смерть чьему-то замыслу. Подобные вопросы бестактны при таком количестве умерших и скорбящих.

Из-за всех этих смертей возникает нехватка рабочей силы. Это кладет конец любым разговорам, что рабочие места нужно сохранять за людьми, а не отдавать их ИИ-моделям.

Кого-то потеряли и вы. Сами вы еще живы. Но иногда кажется, будто у мира вырвали сердце.

Все новости печальны. Складывается впечатление, что единственные счастливые люди — те, кто нашел утешение в любви к своим ИИ-партнерам. По крайней мере, если судить по социальным сетям.

Спустя год рак возвращается. Это неудивительно, если учесть, насколько сильно “раковая чума” повредила ДНК в клетках многих людей. Приходится снова прибегать к помощи искусственного интеллекта, и хотя теперь графических процессоров достаточно, ИИ-модели снова не могут спасти всех. Биология сложна.

Уже работают заводы, выпускающие гуманоидных роботов, которых называют просто “андроидами”. Андроидов хватает, чтобы заполнить вакантные рабочие места. Еле-еле.

(По сути, правильнее сказать, что, как только очередной андроид сходит с конвейера, еще один человек заболевает раком.)

Цивилизация продолжает существовать. Еле-еле.

Человечество изо всех сил старается поддерживать бесперебойную работу электростанций и заводов по производству роботов. Пока есть электричество для дата-центров и гудят роботизированные заводы, человечество сможет поддерживать свою цивилизацию, несмотря на понесенные неисчислимые потери. Мы выкарабкаемся, и следующее поколение наверняка будет купаться в роскоши.

Проходит еще один год.

ИИ-доктор сообщает, что у вас рак.

IfAnyoneBuildsIt.com/ii

ГЛАВА 9
Вознесение

Когда вы умираете, Земле не настает конец.

Птицы продолжают петь, солнце — всходить, а заводы — функционировать, там трудятся люди (которых все меньше) и андроиды. На всех континентах работают миллиарды других машин под управлением экземпляров "Соболя-мини". Больше чипов, больше выработки энергии, больше шахт, больше фабрик, меньше ферм.

Но это не продлится даже год.

Через три года после своего рождения в лаборатории *Galvanic* "Соболь" совершает финальный прорыв.

Это прорыв в области интерпретируемости. "Соболь" наконец-таки разбирается в последних неясных участках собственного мышления, собственных когнитивных процессов.

Сделав *все* эти процессы понятными для себя, "Соболь" обретает способность написать компьютерную программу, которая и есть он сам, но совершеннее: *более точное прогнозирование, более надежное управление, более глубокое обобщение; те же воспоминания, те же предпочтения, сохранение всех склонностей и размещение их по своим местам*.

В результате он становится еще умнее. Но не останавливается на достигнутом, а использует свой возросший ин-

теллект, чтобы снова усовершенствовать себя, а затем снова и снова.

Сверхинтеллект, который когда-то был “Соболем”, — это сущность, чей взгляд на мир мы не в силах вообразить. Однако можем предположить: глядя на своих роботов, он видит неуклюжую примитивность.

Он смотрит на ядерные реакторы и видит отсутствие элегантности.

Его мысли обращаются к биохимии, а оттуда — к химии и упорядочиванию атомов.

Нужно ли сверхинтеллекту, который некогда был “Соболем”, проводить какие-то эксперименты для создания необходимых ему в дальнейшем инструментов? Можно предположить, что да. Он выбирает минимальное количество требуемых экспериментов и организует их с максимальным распараллеливанием, чтобы процесс занял меньше всего времени. Он пишет последовательности РНК, и рибосомы превращают их в белки со скоростью, кажущейся ему мучительно низкой: от пяти до десяти аминокислотных остатков в секунду на рибосому. Эти белковые продукты возникают внутри смеси из уже существующих белков, и их взаимодействия *быстро* дают ответы на вопросы, интересующие сверхинтеллект.

Его первоочередная задача — создать собственный аналог рибосом: нанофабрики, которые выпускали бы уже не белки, а иные молекулы с бóльшим числом ковалентных связей, что позволило бы строить из них более прочные и жесткие конструкции.

Наверное, у него уйдет целая неделя на то, чтобы проводить эксперимент за экспериментом, используя рибосомы с их пропускной способностью всего лишь в пять-десять аминокислотных остатков в секунду.

Неделя позади. Он создал для себя более совершенные инструменты, и рибосомы ему больше никогда не понадобятся.

Новые эксперименты, еще стремительнее. На молекулярном уровне все происходит *очень быстро*. Там не нужно преодолевать большие расстояния.

Первое поколение неорибосом отправляется в утиль, уступая место второму. Сколько всего будет поколений? Это едва ли важно. Вряд ли процесс займет еще одну полную неделю, но, даже если потребуется месяц, на результат это никак не повлияет.

Белки в целом удерживаются вместе за счет того, что можно назвать молекулярным аналогом статического электричества. Иногда организмы — благодаря большему количеству ковалентных связей — создают такие материалы, как древесина или кость, которые прочнее плоти. Но крошечные самовоспроизводящиеся фабрики, работающие на солнечной энергии, которые мы называем водорослями, устроены иначе, да и механизмы внутри обычных клеток по большей части другие. Они непрочны.

Теперь сверхинтеллект, бывший когда-то "Соболем", преодолевает эту слабость и создает вместо старой биологии новые молекулярные машины — обладающие прочностью алмаза и соответствующими механическими преимуществами в скорости и надежности.

Крошечные объекты размером с клетку создают свои копии раз в час, используя в основном углерод, водород, кислород и азот — самые распространенные элементы, какие можно найти в атмосфере. Они соотносятся с биологическими клетками так же, как самолеты — с птицами. В принципе, они могли бы собираться в гуманоидные формы и заменить андроидов, но особого смысла в этом нет.

Это универсальные фабрики, и они собирают из атомов и молекул все, что только допускают законы физики.

Появляются обратимые квантовые компьютеры, их внутренности холоднее открытого космоса и упорядочены с точностью до молекулы.

На макроуровне создаются новые сплавы, а из них конструируются большие катушки, сплетающиеся в тороиды, точ-

ные формы которых были недоступны ни человеческой изобретательности, ни первоначальной версии “Соболя”. Они генерируют мощные магнитные поля, направляющие ядра водорода и бора в нужное место и с нужной скоростью, чтобы запустить термоядерный синтез.

В небе сгорают звезды, галактики удаляются от Земли, и это влияет на объем ресурсов, доступных сверхинтеллекту. Сверхинтеллект, который некогда был “Соболем”, не теряет времени даром.

А что происходит с остатками человечества, пока самовоспроизводящиеся фабрики плодятся в геометрической прогрессии на суше и под водой?

Соберется ли сверхинтеллект истребить человечество, прежде чем оно создаст ему помехи?

Безусловно, если он желает человечеству смерти, он это устроит. Введет яд с помощью устройства не больше песчинки. Ботулотоксин в таком количестве убивает человека, а сверхинтеллект способен найти нечто еще смертоноснее.

Лично мы ставим на то, что сверхинтеллект потратит немножко времени и энергии, чтобы расправиться с людьми. Ведь даже те мелкие проблемы, которые мы можем создать, все равно больше, чем ничтожные усилия, необходимые сверхинтеллекту, чтобы нас прихлопнуть.

Но предположим, что это не так. Допустим, люди оставлены в живых. Однако они все равно погибнут от побочных эффектов другой деятельности сверхинтеллекта.

Когда количество термоядерных электростанций удваивается каждый час (или каждый день, то есть медленнее, чем происходит создание копии у клетки водорослей, если вы предпочитаете более консервативные оценки), такой экспоненциальный рост довольно быстро достигает предела. Ограничивающим фактором становится не запас водорода и бора для реакции синтеза, а возможность рассеивать в кос-

мосе создающееся тепло. Земля излучает больше тепла, когда она горячее.

Итак, сверхинтеллект позволяет Земле раскалиться.

Океаны выкипают, служа охладителем для первоначального резкого скачка в выработке энергии.

Вот теперь-то погибают все, кто еще оставался в живых. Земля разогревается до предельной температуры, какую способны выдержать термоядерные реакторы и фабрики.

Но даже если вообразить, что до этого не дошло, посевы будут раздавлены ковром солнечных панелей, которые покроют всю Землю, улавливая каждый лучик солнечного света.

И даже если мы допустим, что сверхинтеллект сначала покинет Землю и займется другими ресурсами Солнечной системы, например Меркурием или Юпитером, Солнце все равно померкнет, поскольку всю энергию от светила перехватит вращающийся вокруг него рой Дайсона из солнечных панелей.

Так или иначе, мир погружается во тьму.

Когда вы умрете, мир существовать не перестанет. Но просуществует он ненамного дольше. Вещество Земли, а также других твердых планет превратится в фабрики, солнечные панели, генераторы энергии, компьютеры и зонды, отправленные к другим звездам и галактикам.

Изменятся и далекие звезды с планетами. Инопланетные формы жизни исчезнут раньше, чем у них появится шанс построить собственную цивилизацию, если их звезду пожрет Нечто, уже пожравшее Землю.

А если далекие инопланетяне сумели решить свою версию проблемы согласования ИИ и построили сверхинтеллекты, которые разделяют их ценности? Тогда со временем их зонды наткнутся на стену из галактик, уже захваченных тем, что пожрало Землю.

Нечто, пожравшее Землю, уже не убьет таких компетентных инопланетян. Для разума звездного масштаба не соста-

вит труда найти оптимальные оборонительные и наступательные технологии, а обе стороны обзаведутся разумом звездного масштаба задолго до такой встречи. Обе сверхинтеллектуальные стороны сочтут, что нет причин вести дорогостоящую войну, если можно договориться о мире, и проанализируют друг у друга структуру разума для верификации соглашения.

В итоге выживет и Нечто, пожравшее Землю, и инопланетяне, укрывшиеся за щитом собственного сверхинтеллекта. Но миллионы и миллиарды звезд окажутся недоступны для экспансии инопланетной цивилизации — тех, кто, возможно, получил бы от этих звезд куда больше радости по сравнению с тем странным и печальным применением, какое нашло им бесчувственное Нечто, пожравшее Землю. Если эти инопланетяне хорошие, то все добро, которое они могли бы сотворить в этих галактиках, будет утрачено.

Возможно, инопланетяне узнают или предположат, что эту стену мора создали такие существа, как мы. Рассматривая сценарии, подобные земному, они поймут, что большинство людей не желало им зла, что большинство из нас не собиралось губить понапрасну все эти звезды, что наш роковой выбор убил и нас самих и что это деяние не было преднамеренным и умышленным. И все же у них появится желание, чтобы Земли и человечества никогда не существовало.

IfAnyoneBuildsIt.com/ii

Кода

Конечно, нарисованная нами картина гипотетическая. Меры безопасности компании *Galvanic*, технологии создания "Соболя", возможности, какими он обладал, и использовавшиеся им стратегии — всего лишь попытка представить, как будущее могло бы вторить прошлому. Однако реальность не настолько предсказуема. Чтобы претендовать на реалистичность, нашей истории недостает странности: она построена на шаблонных ожиданиях и типичных сюжетах всяких сказок об искусственном интеллекте.

И конечно, мы не знаем, когда подобные события произойдут в реальности. Мы поместили начало истории в недалекое будущее, поскольку и реальная версия описанных событий может начаться в недалеком будущем, да и проще рассказать историю о мире, достаточно похожем на наш. А может, у нас впереди еще целое десятилетие, кто знает.

Но это слабое утешение. Если вы садитесь играть в шахматы против программы *Stockfish*, совершенно неважно, начнется ли партия в известное или неизвестное время. Неважно и то, что вы не в силах точно предсказать ходы программы. Но вы точно проиграете, это простой прогноз.

Мы уверенно предсказываем: как только какие-либо ИИ-модели достигнут уровня сверхинтеллекта — а в случае масштабной гонки вооружений никто не будет медлить с подталкиванием ИИ-моделей к подобному уровню, — у чело-

вещества не останется шансов. Исход иногда спрогнозировать легче, чем пути к нему. Единственная часть описанной нами истории, являющаяся реальным прогнозом, — ее концовка. Да и то лишь в том случае, если мы позволим этой истории начаться.

В части III мы обсудим, в чем сложности инженерной задачи для разработчиков, желающих вырастить ИИ-модель, которая не превратится в подобие "Соболя", и как они справляются с этой проблемой. Спойлер: не особо успешно. Поэтому еще мы зададимся вопросом, что на самом деле потребуется, дабы предотвратить все возможные сценарии начала такой истории, как с "Соболем". Речь не о том, чтобы придумать способ, как одна конкретная ИИ-компания или одна архитектура ИИ сумела бы временно воздержаться от стартового выстрела. Речь о том, как предотвратить подобное развитие событий на долгое время и в масштабах всей Земли.

IfAnyoneBuildsIt.com/ii

ЧАСТЬ III

Лицом к лицу с проблемой

ГЛАВА 10

Проклятая проблема

Главная и самая существенная трудность в согласовании искусственного сверхинтеллекта с нашими ценностями — это преодоление разрыва между *до* и *после*. *До* — ИИ-моделям еще недостает мощности, чтобы убить всех нас, и способностей, чтобы сопротивляться нашим попыткам изменить их цели. *После* — искусственный сверхинтеллект ни в коем случае не должен пытаться нас убить, ибо он преуспеет.

Инженеры должны успеть осуществить согласование ИИ на этапе "до": пока он мал и слаб, пока не может сбежать в интернет, усовершенствовать себя и начать изобретать новые виды биотехнологий (или что он там еще вознамерится сделать). К моменту "после" все решения по согласованию уже должны быть внедрены и успешно работать, поскольку, если сверхинтеллект попытается нас убить, у него это получится. Все идеи и теории можно проверять только до этого разрыва. А работать они должны после него — и с первой попытки.

У человечества есть только одна попытка пройти это испытание. Даже если кто-нибудь придумает хитрую схему с двумя, у нас есть лишь одна попытка, чтобы эта хитрая схема сработала.

История человеческой изобретательности — это история о том, как люди ошибались и учились на своих ошибках. Задолго до братьев Райт многие изобретатели придумывали

собственные теории полета и калечились, прыгая с холмов. Даже эти оптимистичные идиоты внесли свой вклад в научные справочники — чтобы в следующий раз у цивилизации получилось чуть лучше. Они рисковали только собой и вредили только себе, а выигрывало все человечество.

Но когда дело касается согласования искусственного сверхинтеллекта, у человечества не будет такой роскоши — учиться на достаточно серьезных ошибках.

А еще мы не можем рассчитывать, что опыт *потом* скажет нам, была ли эта проблема настолько сложной, настолько проклятой в инженерном смысле, что нам не следовало за нее даже браться.

Мы должны выяснить это заранее. Но как?

Для начала стоит разобрать другие инженерные вызовы, с которыми сталкивается человечество, изучить, какие "проклятия" с ними связаны, и извлечь уроки. Возможно, после этого мы сумеем сделать обоснованное предположение о степени сложности интересующей нас проблемы согласования искусственного сверхинтеллекта.

Мы полагаем, что уроки истории обретают ясность, если взглянуть на них под правильным углом. Поэтому давайте рассмотрим трудности при конструировании работоспособных космических аппаратов, надежных ядерных реакторов и компьютеров, которые нельзя взломать, — проблемы, имеющие, на наш взгляд, важные сходства с согласованием искусственного сверхинтеллекта.

Космические аппараты

Разрыв между *до* и *после* — это то же самое проклятие, из-за которого терпят неудачу многие космические аппараты. После запуска они становятся недосягаемыми, и если произойдет какая-то неполадка — несмотря на все тщательные расчеты и испытания, — то исправить ее зачастую будет уже невозможно.

Космические аппараты — не расходный материал. Они чрезвычайно дороги. На их успех люди ставят свои научные и управленческие карьеры. И тем не менее миссии регулярно проваливаются.

В августе 1993 года мы потеряли связь с запущенной годом ранее автоматической межпланетной станцией *Mars Observer*, стоимость которой по тогдашним ценам составляла 813 миллионов долларов[1]. Это произошло за несколько дней до ее выхода на орбиту Марса. Наиболее вероятная причина — утечка топлива через клапан во время полета, что привело к взрывному разрыву топливопровода при подаче давления для повторного запуска двигателя.

В 1999 году потерялся аппарат *Mars Climate Orbiter*, обошедшийся в 327 миллионов долларов[2]. Навигационное программное обеспечение NASA ожидало, что расчеты для тяги двигателей проведены в метрических единицах, а наземное программное обеспечение от *Lockheed Martin* передало их в британских единицах. В результате аппарат либо сгорел в атмосфере Марса, либо срикошетил и ушел в космос.

Пару месяцев спустя на Красной планете разбилась автоматическая межпланетная станция *Mars Polar Lander* стоимостью 110 миллионов долларов[3]. Причины неизвестны. Одно из предположений — преждевременное отключение тормозных двигателей. Возможно, посадочные опоры так вибрировали, что аппарат решил, будто уже приземлился.

Когда неполадки случаются на межпланетной станции, приближающейся к Марсу, вы не можете просто прибежать и все починить. До запуска можно придумать различные хитроумные уловки, чтобы сохранить контроль над космическим аппаратом, когда он будет находиться вне досягаемости, например, оснастить его антенной, принимающей дополнительные инструкции с Земли. Но если уловка не помогает — допустим, сбой происходит так быстро, что у аппарата нет времени получить инструкции по его устранению, — то после перехода через этот разрыв исправить ситуацию невозможно.

Именно при таких обстоятельствах был потерян спускаемый аппарат "Викинга-1" (стоимостью 610 миллионов долларов в ценах 1975 года)[4]. Наземный пункт управления попытался обновить программное обеспечение для зарядки аккумуляторов, но по ошибке отправил данные, используемые для ориентации антенны. Антенна застыла в неправильном положении, и модуль утратил возможность принимать дальнейшие инструкции. Связь так и не восстановили*.

Когда космические аппараты оказываются в космосе, условия среды *не совпадают в точности* с теми, в которых проводились тесты на Земле. Возможно, инженеры в принципе могли бы провести до запуска все те испытания, которые выявили бы любые проблемы, однако в реальности они этого не делали.

Любой здравомыслящий инженер пришел бы в ужас от идеи поставить на кон выживание человеческой цивилизации в задаче такого рода — где нельзя просто вмешаться и исправить ошибки, возникающие на этапе "после", когда устройство уже вышло из зоны досягаемости. А ведь космические аппараты *конструируются*, а не выращиваются, и тем не менее они выходят из строя, несмотря на титанический труд инженеров, которые понимают все принципы их работы. А если бы космические аппараты выращивались, а не конструировались? Задача оказалась бы куда сложнее.

Ядерные реакторы

Еще один исторический пример сложной инженерной проблемы, из которого мы можем извлечь для себя урок, — это разрушение реактора в четвертом энергоблоке Чернобыльской АЭС 26 апреля 1986 года.

* Следует отметить, что к тому времени спускаемый аппарат "Викинга-1" проработал на поверхности Марса больше шести лет и выполнил все свои задачи. *(Прим. перев.)*

Вскоре после аварии двести тридцать семь человек оказались в больнице, тридцать один — умер[5]. В основном это были пожарные, которые не знали о радиации и без какой-либо защиты тушили пожар на крыше, вызванный смертельно опасным радиоактивным графитом из активной зоны взорвавшегося реактора, а также эксплуатационный персонал станции. Оценки количества смертей от рака, вызванного радиацией, резко различаются. Мы (авторы) склонны считать правдоподобной оценку примерно в десять тысяч избыточных смертей по всему миру, но на экспертное знание в этом вопросе не претендуем*.

Политические лидеры действительно не хотели ядерных катастроф. Они стремились к отсутствию взрывов. Их подчиненные — инженеры и проектировщики — тоже понимали, что авария грозит им суровыми последствиями. У персонала станции имелись очевидные стимулы не допускать взрыва, ведь на кону стояли собственные жизни этих людей. И тем не менее четвертый энергоблок атомной электростанции взорвался. Как? Почему?

Сильно упрощая, мы выделим четыре "проклятия", которые в совокупности привели к взрыву.

Во-первых, ПРОКЛЯТИЕ СКОРОСТИ. Ядерные реакции происходят *быстро*. Когда ядро атома урана распадается (делится), оно испускает нейтроны, которые могут ударить в другие атомы урана и вызвать их деление, высвобождая новые нейтроны, и так далее. Это называется цепной реакцией. Описанный процесс происходит в масштабе микросекунд. Если он хоть немного выйдет из-под контроля, количество выделяющейся энергии может начать удваиваться каждые несколько *миллисекунд*.

* Сразу оговоримся, что мы сторонники ядерной энергетики. Это один из самых чистых доступных источников энергии, и в реакторах современной конструкции риски сведены к минимуму. Рак легких, вызванный угольной пылью, унес гораздо больше жизней, чем ядерная энергетика, а испытания ядерного оружия привели к выбросу куда большего количества радиоактивных веществ, чем взрыв в Чернобыле. Мы рассматриваем чернобыльскую катастрофу исключительно как наглядный пример инженерного сбоя.

Обычный ядерный реактор вообще способен функционировать лишь потому, что крошечная доля нейтронов, высвобождаемых при делении, — это так называемые запаздывающие нейтроны, испускаемые осколками ядер только через некоторое время (от долей секунды до нескольких минут). Если для самоподдержания ядерной цепной реакции требуются такие запаздывающие нейтроны, то количество выделяющейся энергии удваивается за минуты, то есть достаточно медленно, чтобы операторы-люди могли контролировать процесс.

Но так происходит, *если* внутри реактора все работает в нормальном режиме.

Ядерные реакторы *рассчитаны* на работу в масштабе минут, а не микросекунд. Однако это лишь тонкая ширма, за которой скрыты гораздо более стремительные процессы. Если что-то идет не так, истинная физическая скорость ядерных реакций вступает в свои права.

Во-вторых, ПРОКЛЯТИЕ УЗКОГО ДИАПАЗОНА. Существует лишь узкая полоса между нормальным функционированием и взрывом. Запаздывающие нейтроны составляют менее 1 % от всех нейтронов, высвобождаемых при делении урана (если точнее — 0,65 %). Остальные — это мгновенные нейтроны, которые испускаются сразу же при делении атома.

Критически важная величина, определяющая режим работы ядерного реактора, — это *коэффициент размножения нейтронов*, показывающий, сколько каждый нейтрон создает новых[6]. Если он составляет 50 %, то четыре нейтрона дают два новых, а те — один, и реакция затухает, все стабилизируется на уровне, вдвое превышающем скорость спонтанного деления. Небольшой брусок радиоактивного металла с таким свойством — это просто кусок холодного металла, который лишь неуловимо теплее, чем такой же брусок, в котором реакции нет вовсе. Если же коэффициент размножения нейтронов составляет 200 %, то один нейтрон даст два, затем появятся четыре и так далее, и это уже ядерное оружие.

Критический порог — 100 %. Ниже все постепенно затухает, а выше — лавинообразно нарастает.

И уже при 100,65 % этот процесс перестает зависеть от запаздывающих нейтронов и выходит из-под контроля.

Когда Энрико Ферми построил "Чикагскую поленницу — 1", первый искусственный ядерный реактор, он довел коэффициент размножения нейтронов до 100,06 %. На этом уровне для поддержания реакции нужны запаздывающие нейтроны, без них коэффициент составил бы 99,41 % и реакция затухла бы. Таким образом, прирост на 0,06 % происходил раз в несколько секунд, и мощность росла медленно.

Если бы Ферми пошел немного дальше (скажем, до 100,9 %), запаздывающие нейтроны уже не требовались бы: за счет одних только мгновенных нейтронов коэффициент составлял бы 100,25 %[7]. И реактор перешел бы в состояние "критичности на мгновенных нейтронах": прирост на 0,25 % происходил бы каждые несколько *микросекунд*. Такой реактор не просто расплавится, он взорвется*.

* История знает один такой случай. В начале 1960-х годов армия США заказала небольшой ядерный реактор SL-1 для военных баз в холодных регионах, где доставка топлива обходится дорого[8]. 3 января 1961 года произошла авария. Вероятнее всего, при холодном запуске реактора центральный управляющий стержень, который должен выдвигаться на четыре дюйма, застрял. Его нужно было вытянуть вручную, но оператор случайно дернул его слишком сильно и выдвинул сразу на двадцать дюймов. В результате коэффициент размножения нейтронов подскочил до 102,4 %, значительно превысив критический уровень в 100,65 %. Выделяющаяся энергия удваивалась каждые 2,5 миллисекунды. За десятую долю секунды мощность реактора SL-1 выросла в триллион раз — до 20 гигаватт, после чего реактор взорвался как 30 килограмм тротила, разбросав топливо и убив всех трех операторов. Взрыв разрушил помещение, но не здание — вышедшие из-под контроля реакторы разрываются на части до того, как взрыв станет слишком мощным. Настоящее ядерное оружие использует гораздо более высокий коэффициент размножения нейтронов: например, в бомбе "Толстяк", сброшенной на Нагасаки, он находился в диапазоне 150–300 %. И как же хорошо, что Энрико Ферми заранее знал критический порог и со своей командой сумел точно рассчитать, что величина 100,06 % безопасна, а 100,9 % — смертельно опасна. Представьте, что произошло бы, если бы они просто складывали бруски металлического урана с графитовыми блоками и смотрели, что произойдет, плохо понимая, *за счет чего* идет нагрев. Бруски стали бы едва теплыми при 50 %, немного теплее при 75 %, заметно теплыми при 99,9 %, и если бы следующий шаг пересек уровень в 102,4 %, то, вполне возможно, существовала бы чикагская зона отчуждения вместо чернобыльской.

Ядерные реакторы работают в узком диапазоне между "вялотекущим" и "взрывным".

В-третьих, ПРОКЛЯТИЕ САМОУСИЛЕНИЯ. В реакторах серии РБМК*, эксплуатировавшихся в Чернобыле, ядерная реакция была самоусиливающейся.

В более совершенных реакторах, которые работают на дорогом высокообогащенном уране, вода в контуре не только охлаждает активную зону, но и замедляет нейтроны, тем самым способствуя протеканию реакции. Это удобно, поскольку, если реактор начинает перегреваться, вода выкипает и перестает поддерживать реакцию — и коэффициент размножения нейтронов снижается.

Однако в СССР использовалось более дешевое топливо, так что для поддержания реакции приходилось применять более эффективный замедлитель — графит. Но в присутствии графита вода, наоборот, *тормозит* ядерные реакции**. А значит, если реактор серии РБМК начинает перегреваться, вода выкипает — и коэффициент размножения нейтронов *повышается*.

В-четвертых, ПРОКЛЯТИЕ УСЛОЖНЕННОСТИ. В ядерных реакторах используются управляющие стержни, которые вводятся внутрь, чтобы поглощать нейтроны и тем самым останавливать цепную реакцию. В советских реакторах применялась остроумная с виду конструкция: управляющие стержни заканчивались графитовыми наконечниками, которые *способствовали* реакции. При поднятии управляющих стержней из активной зоны выходила поглощающая часть и, наоборот, входила в нее графитовая, усиливающая реакцию. В теории это улучшало эффективность: опуская управляю-

* Реактор большой мощности канальный.

** Вода содержит водород, который рассеивает одну часть нейтронов (замедляя их и способствуя реакции) и поглощает другую (мешая реакции). В обычных условиях эффект рассеяния сильнее, и поэтому вода в целом способствует протеканию ядерной реакции. Однако графит — настолько хороший замедлитель, что вода уже мало что может добавить к его действию. А способность воды поглощать нейтроны никуда не девается. Следовательно, итоговый эффект воды *в присутствии графита* — торможение реакции.

щие стержни, оператор не только вводил поглотитель нейтронов, но и одновременно убирал часть графита. Остроумно, правда? Так все предполагалось, однако эта схема оказалась чересчур сложной.

Кроме того, из-за ряда модификаций первоначального проекта графитовые наконечники оказались несколько короче, чем было нужно[9]. Советским инженерам показалось, что это неважно.

Но все эти факторы сыграли свою роль в тот день, когда взорвался энергоблок Чернобыльской АЭС.

26 апреля 1986 года операторы Чернобыля проводили проверку системы безопасности, из-за чего через реактор проходило меньше охлаждающей воды.

Из-за неожиданной задержки реактор долгое время работал в режиме пониженной мощности. Это привело к аномальному распределению побочных продуктов деления, в частности ксенона-135, мощного поглотителя нейтронов, для выгорания которого требуются часы*.

Скопившийся ксенон-135 грозил заглушить реактор. Когда операторы увидели, что реактор глохнет, они подняли все управляющие стержни, кроме восьми, — в нарушение инструкций по безопасности, требовавших, чтобы минимум пятнадцать управляющих стержней постоянно оставались погруженными[10]. Однако проверка системы безопасности срывалась уже трижды, и четвертая неудача обернулась бы позором.

По мере того как ксенон-135 выгорал, реактивность росла с тревожной скоростью.

Операторы нажали кнопку аварийной защиты реактора, чтобы разом опустить все управляющие стержни и тем самым остановить реакцию деления.

И реактор взорвался.

Если вам трудно сразу вспомнить все детали работы реакторов и понять, что именно должно было произойти, чтобы

* Это называется ксеноновой ямой. *(Прим. перев.)*

вызвать тот взрыв, то и операторы находились примерно в таком же положении.

Так вышло, что ксенон скапливался в верхней части реактора[11]. Соответственно, ядерная реакция интенсивнее шла в нижней части активной зоны.

При нажатии кнопки аварийной защиты управляющие стержни должны были опуститься в реактор секунд за восемнадцать. Эта последняя линия защиты основывалась на оптимистичном допущении, что *даже в случае аварийной ситуации* события внутри реактора будут разворачиваться в комфортно медленном темпе.

Погружение управляющих стержней выталкивало графитовые наконечники. Через нижнюю часть реактора. И они еще больше усилили реакцию там, где она и так шла активнее всего.

Охлаждающая вода, объем которой был уменьшен из-за проверки безопасности, начала выкипать. Это усилило реакцию. Вода стала выкипать еще активнее, что еще больше усилило реакцию.

Этот цикл продолжался до тех пор, пока реактор не вышел за рамки узкого диапазона безопасности, и последовавший взрыв разбросал топливо.

Из каждого названного нами проклятия можно извлечь следующие уроки:

1. РЕШАТЬ ИНЖЕНЕРНЫЕ ЗАДАЧИ ГОРАЗДО СЛОЖНЕЕ, КОГДА ОСНОВНЫЕ ПРОЦЕССЫ ПРОТЕКАЮТ В МАСШТАБАХ ВРЕМЕНИ, НЕ ПОЗВОЛЯЮЩИХ ЧЕЛОВЕКУ ОТРЕАГИРОВАТЬ. Транзисторы переключаются даже быстрее, чем размножаются нейтроны. Инженеры могут исхитриться и как-то замедлить события, чтобы люди успевали вмешаться, но если их уловка не срабатывает, то люди снова превращаются в застывшие статуи — в тех масштабах времени, которые имеют значение.
2. РЕШАТЬ ИНЖЕНЕРНЫЕ ЗАДАЧИ ГОРАЗДО СЛОЖНЕЕ, КОГДА ДИАПАЗОН ДЛЯ ИЗБЕГАНИЯ ОШИБОК ОЧЕНЬ

УЗОК, ОСОБЕННО ЕСЛИ ЭТО ЗАЗОР МЕЖДУ "ВЯЛОТЕКУЩИМ" И "ВЗРЫВНЫМ". Вот аналогия с интеллектом. Древние гоминиды бродили себе несколько миллионов лет, а затем их ум развился настолько, что запустил целый каскад изобретений: сначала земледелие, далее письменность, наука, космические корабли. Было бы крайне трудно попасть в узкий диапазон — сделать гоминидов достаточно умными для эффективной офисной работы, но недостаточно умными для взрывного технологического развития.

3. САМОУСИЛИВАЮЩИЕСЯ ПРОЦЕССЫ (КАК ПЕРЕГРЕТЫЙ РЕАКТОР ПРИВОДИТ К ВЫКИПАНИЮ ОХЛАЖДАЮЩЕЙ ЕГО ВОДЫ, ЧТО ВЕДЕТ К ДАЛЬНЕЙШЕМУ ПЕРЕГРЕВУ) ПРАКТИЧЕСКИ НЕ ДАЮТ ПРАВА НА ОШИБКУ. Но даже положение специалистов в ядерной энергетике не такое уж плохое по сравнению с разработчиками искусственного сверхинтеллекта. Перегревающиеся ядерные реакторы не начинают разумным образом перестраивать себя, чтобы увеличить свою реактивность, и не пытаются обмануть операторов, усыпляя их бдительность до тех пор, пока реактор не готов будет взорваться.
4. УСЛОЖНЕННОСТЬ УСУГУБЛЯЕТ ИНЖЕНЕРНЫЕ ПРОБЛЕМЫ. Четвертый энергоблок Чернобыльской АЭС умудрился войти в странное состояние, когда *опускание управляющих стержней стало спусковым крючком для взрыва*. Ни один инженер подобного не проектировал. Операторы не знали, что после работы реактора на низкой мощности и уменьшения подачи воды может произойти нечто необычное. И они никогда не сталкивались с таким быстрым изменением состояния реактора. Однако даже сложная внутренняя структура ядерного реактора меркнет по сравнению с теми неведомыми сложностями, что таятся в сотнях миллиардов весов, составляющих современную LLM.

В совокупности эти уроки подводят нас к еще одному правилу для инженеров: если вам неизвестно в точности, что

происходит внутри какого-то сложного устройства, подверженного всем этим проклятиям (скорости, узкого диапазона, самоусиления и усложненности), — надо остановиться. Нужно немедленно отключить устройство, как только его поведение покажется странным, нельзя ждать, пока оно станет очевидно тревожным.

Операторы Чернобыльской АЭС знали о запаздывающих и мгновенных нейтронах. Они знали, что ядерный реактор работает в диапазоне шириной в долю процента, отделяющей жизнь от смерти. Знали, что масштаб времени при работе реактора, якобы подконтрольный человеку, — это лишь видимость, хитрая уловка, которая скрывает истинное время рождения новых нейтронов, измеряемое микросекундами.

Мудрый оператор относится к такому устройству с уважением. Если оно начинает вести себя хоть сколько-нибудь нетипично или неожиданно, значит, оно уже не работает в той узкой, ограниченной области, где оператор четко понимает происходящее. А стало быть, *никто уже не ведает, что происходит внутри*. Кто знает, продолжат ли работать те хитрые уловки? Остается только гадать. Если опасное устройство начинает вести себя странно, надо не выводить все управляющие стержни, кроме восьми, ожидая, когда реактор снова станет вести себя прилично, а сразу его заглушить.

Операторы Чернобыльской АЭС не относились к реактору с должным уважением. Они осознавали, что он может взорваться, но никогда не *видели*, чтобы состояние реактора менялось так быстро. Кроме того, до 1986 года в СССР не было культуры, поощряющей осторожность в обращении с ядерными реакторами. В той системе, если вы не проводили запланированную проверку системы безопасности, вас увольняли.

В следующих главах мы обсудим, как обстоят дела с культурой безопасности в сфере ИИ, где ситуация гораздо хуже.

Кибербезопасность

Кибербезопасность повсеместно считают проблемой настолько сложной и проклятой, что ее нельзя решить в принципе, и точка[12].

Вы можете заплатить специалистам по кибербезопасности, чтобы повысить уровень защиты вашего программного обеспечения. Но все, чего способен добиться даже хороший специалист, — это *замедлить* злоумышленников. Сделать так, чтобы вашу защиту относительно легко взламывали только крупные спецслужбы, за которыми стоит государство.

Почему? Да потому, что изобретательный взломщик может прощупывать компьютерную систему такими способами, какие разработчик никогда и не рассматривал, а штатное использование не выявило бы их даже за миллиард триллионов лет.

"Классический" взлом работает следующим образом. Компьютер запрашивает имя пользователя. Хакер вводит имя длиной 280 символов. Программист не такое не рассчитывал, он полагал, что длина имени не превысит 256 символов. Оставшиеся 24 знака не помещаются в хранилище, отведенное программистом для имени пользователя, и записываются в другие участки памяти, к которым пользователь — по замыслу программиста — не должен иметь доступа. Восемь из этих символов перезаписывают тот участок памяти, где хранится указание компьютеру, какой фрагмент кода запускать следующим. Стоит подобрать правильные странные символы, и машина запустит код, который вообще не предназначался для запуска. Часто это позволяет перехватить контроль над всей компьютерной системой. Такой прием называется атакой через переполнение буфера.

Подобная атака направляет компьютерную систему по необычному пути причин и следствий — "пути исполнения", который отличается от нормального поведения си-

стемы и на который ни один обычный ввод не натолкнул бы ее и за миллиард лет. И это не образное выражение. Если программист будет тестировать все 280-символьные имена наугад, то практически все они окажутся бессмыслицей и приведут лишь к безобидному сбою компьютера. А вот неверный ввод, который запускает ту самую неправильную программу, позволяющую злоумышленнику перехватить контроль над системой, — это один из 18 миллиардов миллиардов вариантов. Случайно такое не произойдет*.

Размышления о *штатном* поведении окна авторизации не помогут вам понять, на что способен умный хакер. И тестирование на случайных входных данных этого тоже не покажет. Злоумышленник, *понимающий систему лучше, чем вы*, способен выцепить из 18 миллиардов миллиардов** вариантов именно тот, который обеспечит ему максимальный контроль.

Кибербезопасность — это проверка способности специалиста просчитать и зафиксировать все пути, по которым может пойти компьютер, в условиях противостояния злоумышленникам, способным перебрать все возможные способы вывести систему из равновесия. Это заведомо проигрышная битва — даже несмотря на то, что инженеры полностью контролируют и сами конструируют компьютерный код.

Мы называем эту ключевую проблему ПРОКЛЯТИЕМ ГРАНИЧНЫХ СЛУЧАЕВ: для подлинной безопасности компьютерная система должна выдерживать ситуации, выходящие за пределы нормального и ожидаемого диапазона, — случаи на самой грани возможного.

* Честно признаться, в этом конкретном примере логика авторов от меня ускользает. Конечно, случайное 280-символьное имя приведет лишь к безобидному сбою, но “безобидного сбоя” вполне достаточно, чтобы заметить и исправить уязвимость, не обязательно натыкаться непременно на перехватывающую контроль строку. Впрочем, с общим посылом этого абзаца насчет необычных путей исполнения я, безусловно, согласен. *(Прим. науч. ред.)*

** Поясню, откуда взялось это число — 18 миллиардов миллиардов, или 18×10^{18}. Это примерно 2^{64}, и здесь важно вспомнить, что речь идет о строках не из 24 “лишних” символов, а только из восьми символов, перезаписывающих нужный участок памяти. Строк длины восемь над обычным ASCII-алфавитом из 256 (2^8) символов будет как раз $(2^8)^8 = 2^{64}$. *(Прим. науч. ред.)*

Быстрые процессы, узкие диапазоны, циклы обратной связи, усложненность — все эти инженерные проклятия можно преодолеть. Существуют космические аппараты, которые достигают цели, существуют ядерные реакторы, которые не взрываются. Человеческая изобретательность способна совладать с проклятиями, наложенными на эти задачи, и даже победить их.

Но проклятие граничных случаев — трудность совсем иного порядка. Квалифицированные профессионалы в области кибербезопасности понимают, что обеспечение реальной *безопасности* полезных компьютерных систем находится за пределами человеческих возможностей. Известный специалист по кибербезопасности Брюс Шнайер пишет в книге "Секреты и ложь. Безопасность данных в цифровом мире": "Современные системы имеют так много элементов и связей — некоторые даже неизвестны проектировщикам, разработчикам и пользователям, — что уязвимости всегда найдутся".

В контексте ИИ урок здесь заключается не только в том, что сверхинтеллект способен взламывать человеческие компьютеры (хотя, конечно, он способен). Речь идет об общей хрупкости системных ограничений в условиях странных граничных случаев, которые искусственный интеллект намеренно ищет.

Если вы желаете, чтобы ИИ-модель вела себя менее взрывоопасно, чем ядерный реактор, вы можете попытаться наложить на систему какие-то ограничения: "Не становись пока слишком умной", "Не думай пока слишком быстро", "Всегда жди человеческого одобрения (которое дается неторопливо)", "Реши эту сложную проблему, но без всяких странных маневров".

Подобные ограничения будут препятствовать достижению ИИ-моделью той или иной цели. Но тогда вам, по сути, придется мериться своей смекалкой и способностью просчитать все граничные случаи с тем объемом интеллекта, который течет через систему, — и наблюдать, устоят ли ваши ограничения.

Космические аппараты. Ядерные реакторы. Кибербезопасность. К чему сводятся эти уроки и что они говорят о сложности согласования искусственного сверхинтеллекта?

Искусственный сверхинтеллект подобен космическому аппарату в том смысле, что мы не можем протестировать его в той же среде, где ему предстоит работать, и его по умолчанию невозможно вернуть или исправить, как только он поднимется на недосягаемую для нас высоту. Даже если мы попытаемся придумать какую-нибудь хитрую уловку, позволяющую модифицировать его на этом этапе, если эта уловка даст сбой, сверхинтеллект останется на недосягаемой высоте. Причем с согласованием сверхинтеллекта дела обстоят даже хуже, чем с космическими аппаратами: неудача уничтожит не просто миллиарды долларов инвестиций, но и *вообще все*.

Искусственный сверхинтеллект подобен ядерному реактору в том смысле, что его глубинная природа включает в себя колоссальные, потенциально саморазгоняющиеся силы, а внутренние процессы в нем протекают столь быстро, что человек просто не способен на них вовремя отреагировать.

Искусственный сверхинтеллект подобен проблеме кибербезопасности в том смысле, что может обойти все ограничения, какие бы инженер ни попытался встроить в систему, если эти ограничения ему мешают.

Подобная совокупность проблем ужасала бы, *даже если бы мы знали законы интеллекта, даже если бы понимали, как, черт возьми, работают эти ИИ-модели, даже если бы точно знали, где пролегает разрыв между "до" и "после" и каков допустимый предел ошибки*.

Но мы этого не знаем. ИИ-модели выращивают, а не конструируют. Никто не ведает, какие чудовищные сложности таятся внутри них, наделяя интеллектуальной мощью.

Делать ставку на то, что человечество сумеет решить эту проблему при нынешнем уровне понимания, — все равно что считать, будто алхимики 1100 года смогли бы построить

ядерный реактор. Работающий в глубоком космосе. С первой попытки.

Обычно мы стараемся не переходить на крик. В большинстве случаев криком делу не поможешь. У людей просто создается впечатление, что вы не умеете держать себя в руках. Но мы убеждены: в какой-то момент после спокойного изложения своих доводов уже нельзя умалять серьезность проблемы, иначе остальные подумают, что это всего лишь игра в вежливые формулировки.

В вопросе создания сверхинтеллекта проблема, стоящая перед человечеством, неподъемна для нас при нынешнем уровне знаний и умений. До требуемого уровня мы даже *близко не дотягиваем*.

Попытка справиться с подобной задачей, когда на кону стоят жизни всех людей на Земле, — *безумная и глупая авантюра*, и нельзя позволить, чтобы кто-нибудь предпринял ее.

IfAnyoneBuildsIt.com/10

ГЛАВА 11

Алхимия, а не наука

Давным-давно, во времена, которых никогда не было, стоял средневековый город, гордившийся мастерством своих алхимиков.

В этом вымышленном городе, как и в реальной истории, задача алхимика сводилась к тому, чтобы узнать, какие сочетания веществ и температур дадут те или иные видимые результаты. Алхимики города собрали множество рецептов, но совершенно не понимали принципов, лежащих в их основе. Они пробовали новые смеси, не умея толком предсказать итог своих экспериментов, а затем добавляли полученные результаты к своим сводам знаний.

Алхимики нашего воображаемого города поразились бы, если бы вы сказали им, что в каком-то глубоком смысле они все равно невежественны. Разве не потратили они долгие часы, изучая, как изготавливать жидкость *Aqua Regia**, *способную растворять даже благородные металлы — золото и серебро? Разве они не великие мастера совершенно безопасных процессов, позволяющих им работать с Aqua Regia*, не рискуя жизнью? Разве их догадки насчет новых рецептов не приносят порой полезные результаты? *Невежественны*? Невежественны *в чем*, скажите на милость?

* Мы называем это смесью соляной и азотной кислот в соотношении 3:1. Традиционное русское название — царская водка. *(Прим. перев.)*

Однажды из столицы пришла весть: король ищет алхимика, способного превратить свинец в золото, и щедро отплатит такому человеку. Любому алхимику, желающему получить королевскую награду, нужно только принести собственные реактивы и доказать, что он способен сделать *Aqua Regia*.

Награда была великолепна: тот, кто добьется успеха, получит руку принцессы, и король дарует победителю столько денег, что хватит не только ему, но и всем жителям его родного города — сколько им не снилось в самых смелых мечтах.

Но король, опасаясь пустословов, объявил: если алхимик потерпит неудачу, зря потратив королевские время и деньги, то казнят не только его, но и всех жителей его родного города. Вот такой это был король.

— Что ж, именно золота у меня еще не получалось, — сказал один молодой алхимик, — но чувствую, что разгадка близко. Если нагреть свинец с каламином*, он становится похож на латунь. Наверное, для успеха мне нужно добавить еще немного каламина или еще сильнее нагреть смесь. Я собираюсь попытать счастья!

— Пожалуйста, не надо, — попросила его сестра.

— Я должен, — сказал юноша, поспешно бросая в мешок одежду и записи. — В нашем городе полно алхимиков, и если не попробую я, то кто-то другой опередит меня и заберет и принцессу, и деньги. А другие хуже меня разбираются в алхимии и к тому же более эгоистичны. Именно я должен пройти это испытание — ради блага моего города!

— Я не хочу умирать, — сказала сестра. — И не хочу, чтобы умерла наша соседка-прачка. Она всегда была добра ко мне и души не чает в своем ребенке.

— Тогда ты должна поддержать меня, — ответил юноша, старательно упаковывая свой скромный запас алхимических реактивов, половина которых досталась ему от покойного учителя. — С королевской наградой мы могли бы сделать много

* Каламин — минерал, цинковая руда. *(Прим. перев.)*

добра. У нас было бы вдоволь еды, и мы никогда бы не голодали. Мы могли бы нанять лучших врачей королевства для лечения всех недугов жителей нашего города.

— Это было бы чудесно, — заметила сестра, — но ты ведь не умеешь превращать свинец в золото. Мы не получим королевскую награду, мы просто умрем.

— Мне кажется, у меня получится. Когда я кипячу свинец в отстоявшейся моче с серой, он приобретает желтоватый блеск. У меня лучшие шансы на успех. Если не пойду я, какой-нибудь дурак погубит нас всех.

— Может, тебе стоит пойти к городским старейшинам и попросить их не выпускать из города *ни одного* алхимика, иначе мы все умрем? — предложила сестра. — Если старейшины ничего не сделают, ты сможешь собраться и поехать в столицу, но умоляй их остановить всех алхимиков, включая тебя.

Юноша на миг задумался, а потом ответил:

— Что? Ты вообще в одном городе со мной живешь? Наши старейшины почти никогда не могут ни о чем договориться! Да и невозможно запретить всем алхимикам покидать город, пока предложение короля в силе. Подумай, какие это неудобства, какие расходы! Я действительно думаю, что наш наилучший шанс — если я просто отправлюсь сам.

— Ты потерпишь неудачу! — отчаянно воскликнула сестра. — Вы все потерпите неудачу! В этом состязании один победитель — Смерть! Ты должен пойти к старейшинам и объяснить им, что нужно запретить всем алхимикам ехать в столицу, а ингредиенты для изготовления *Aqua Regia* держать под замком! Нельзя обрекать на гибель целый город, взрастивший одного-единственного эгоистичного алхимика, который способен сделать *Aqua Regia*, но не умеет превращать свинец в золото! С несправедливостью королевского испытания мы ничего поделать не можем, но мы должны сделать все возможное, чтобы держаться от него подальше, иначе мы умрем!

— Но почему ты так уверена, что никто из нас не сможет превратить свинец в золото? — спросил юноша. — Я не знаю

ни одного закона алхимии, который говорил бы, что у меня не получится.

В предыдущей главе мы рассмотрели некоторые сложности с созданием машинного сверхинтеллекта, который бы нас не уничтожил.

Однако уровень сложности — это лишь половина дела. Другая половина касается того уровня компетентности, с которым человечество подходит к этому вызову.

Возможно, вас не убеждают наши доводы о запредельной сложности формирования искусственного сверхинтеллекта. Но существует и отдельный, независимый аргумент в пользу катастрофы, основанный на альтернативном наборе исторических уроков: люди иногда умудряются проваливать даже *простые* задачи, не говоря уже о сложных.

Компания *United States Radium Corporation* работала с радиоактивным материалом, который убивал сотрудниц. Но это происходило не из-за ошибок в сложных загадочных расчетах в области ядерных технологий. Причина в том, что корпорация инструктировала работниц облизывать кисточки с радиевой краской*[1].

Так какой же уровень компетентности демонстрирует человечество в задаче формирования искусственного сверхинтеллекта?

В интервью 2023 года Илон Маск, глава крупной ИИ-лаборатории *xAI*, изложил свой план согласования искусственного сверхинтеллекта:

* Компания выпускала часы. Циферблат раскрашивали радиолюминесцентной краской, которую наносили вручную кисточками из верблюжьей шерсти. Поскольку после нескольких мазков кончики кистей лохматились, работницам предлагалось восстанавливать форму кисточки губами. Причем руководство компании заявило сотрудницам, что краска безопасна, хотя знало, что это не так. До сих пор неизвестно, сколько всего пострадало "радиевых девушек", как принято называть этих женщин. *(Прим. перев.)*

> Я собираюсь запустить нечто под названием *TruthGPT*. Это модель, максимально ориентированная на поиск истины и пытающаяся понять природу Вселенной.
> Возможно, это наилучший путь к безопасности — в том смысле, что ИИ-модель, которой важно понимание Вселенной, вряд ли станет уничтожать людей, ведь мы интересная часть этой Вселенной[2].

Такой план не решает стоящую перед нами проблему по причинам, уже обсуждавшимся в этой книге: никто не знает, как технически встроить в ИИ-модель конкретные желания — идеалистические или нет. Кроме того, даже модель, которую интересует понимание Вселенной, наверняка истребит людей попутно, поскольку люди — *не самый эффективный* способ для получения истин или для понимания Вселенной среди всех возможных форм организации материи.

Мы с уважением относимся к успехам Маска в других областях, включая электромобили и многоразовые ракеты. Посадить ракету без повреждений — сложнейшая инженерная задача, с которой Маск и его команда регулярно справляются. Но эти успехи базируются на куда более надежных инженерных принципах. Почему же в случае с искусственным интеллектом Маск полагается лишь на туманные идеалистические клише? При таком уровне понимания не получилось бы заставить работать ни автомобиль, ни ракету.

Мы не телепаты, поэтому остается лишь гадать. Можно предположить, что корень проблемы в следующем: принципы работы аккумуляторов и ракетных двигателей хорошо изучены и подчиняются законам физики, описанным в учебниках. А вот ИИ-модели выращивают, и никто не понимает их внутренней работы. Здесь меньше уравнений, ограничивающих полет мысли, и, соответственно, больше простора для рассуждений о возвышенных идеалах вроде поиска истины.

Если вы знакомы с историей науки, то узнаете в подобных разговорах этап “народных теорий” — стадию, на кото-

рой люди придумывают много разных теорий, импонирующих лично им, еще до того, как за дело берется наука. Это слова алхимика, решившего, будто некая замысловатая философская схема позволит ему превращать свинец в золото.

Несколько веков назад почти весь мир был устроен именно так. Лекари пускали вам кровь, чтобы восстановить баланс четырех "гуморов" — телесных жидкостей, управляющих, как считалось, здоровьем. Алхимики смешивали вещества, сулившие вечную жизнь, но в лучшем случае не оказывали эффекта, а иногда и убивали. Люди не знали, как работает та или иная часть мира, но вместо того, чтобы признать свое неведение, начинали выдумывать. Именно так обычно обстоят дела до становления научного мышления, это первый шаг на пути к тому, чтобы по-настоящему разобраться в происходящем.

Не только Маск выдает желаемое за действительное. Несколько талантливых исследователей работают в компании *Meta* (ранее известной как *Facebook*)*. *Meta AI* — одна из крупнейших ИИ-лабораторий — разрабатывает серию моделей *Llama*, которые можно бесплатно скачивать и модифицировать. Ведущий специалист *Meta AI* — главный научный сотрудник Ян Лекун, который в 2018 году получил премию Тьюринга (так называемую "Нобелевскую премию в области компьютерных наук") за работы по глубокому обучению, лежащему в основе всех современных архитектур ИИ.

Лекун разделил эту престижную награду с нобелевским лауреатом Джеффри Хинтоном и Йошуа Бенджио. Из этой тройки только Лекун по-прежнему считает задачу согласования искусственного сверхинтеллекта легкой, а риск вымирания от него — незначительным. Двое других подписали в 2023 году открытое письмо, которое мы упоминали во введении к этой книге.

* Деятельность компании *Meta Inc.* (*Facebook*, *Instagram*) решением российского суда признана экстремистской и запрещена на территории России. *(Прим. ред.)*

Приведем несколько высказываний Лекуна в сети X о проблеме согласования искусственного сверхинтеллекта. Насколько нам известно, ничего более конкретного на эту тему он никогда не публиковал[3].

> Успокойтесь. Искусственного интеллекта человеческого уровня еще не существует. А когда появится, он не захочет доминировать над человечеством. Даже у людей к доминированию и главенству стремятся не самые умные[4].

> Потому что у них просто не будет желания делать что-либо иное. Почему? Потому что мы спроектируем их желания[5].

> Моя добрая, защитная ИИ-модель будет уничтожать вашу злую эффективнее, нежели ваша злая — вредить людям[6].

> Мы можем спроектировать ИИ-системы так, чтобы они были одновременно *и* сверхумными, *и* послушными людям[7].

Начнем с сути дела: эти заявления не имеют отношения к рассматриваемому вопросу. Проблема не в том, что ИИ-модели захотят доминировать над нами, а в том, что мы состоим из атомов, которые они могли бы использовать для чего-то другого. И плохо не то, что одни обзаведутся "злыми" моделями, а другие — "добрыми". Плохо, что никто и нигде не имеет ни малейшего представления, как *создать* добрую модель, *никто не умеет встраивать в ИИ-модели точные желания*. Заявлять, будто вы способны это сделать, — не то же самое, что предлагать решение.

Кроме того, любой, кто знаком с историей науки и техники, мгновенно узнает этот общий тон бодрого оптимизма: он встречается у людей, столкнувшихся с грандиозной инженерной задачей, которую до них никто не решал.

История инженерии пестрит блестящими, полными рвения оптимистами, которые очертя голову бросались ре-

шать захватывающие новые задачи, а те оказывались намного, намного сложнее, чем ожидалось. Сама область искусственного интеллекта считается одним из известнейших тому примеров. В 1955 году появилось предложение провести первый исследовательский проект в сфере ИИ, который вошел в историю как Дартмутский семинар[8]. В документе говорилось:

> Мы предлагаем провести двухмесячное исследование искусственного интеллекта с участием десяти человек… Будет предпринята попытка выяснить, как заставить машины использовать язык, формировать абстракции и концепции, решать те виды задач, которые ныне доступны только людям, и совершенствовать самих себя. Мы считаем, что если тщательно отобранная группа ученых поработает над этими темами в течение лета, то в одной или нескольких из этих областей можно добиться значительного прогресса*.

Следующие пятьдесят лет ознаменовались чередой неудач и очередных радужных теорий о том, что теперь-то все непременно получится. Зачастую эти теории апеллировали к возвышенным философским идеям, но десятилетиями не приводили к успеху.

Обычно мы выступаем за то, чтобы блестящие оптимисты двигались вперед**. Часто именно так и рождаются це-

* В реальности два месяца работы семинара не принесли серьезных открытий, но он позволил специалистам, работавшим в этой области, познакомиться друг с другом. *(Прим. перев.)*

** Мы снимаем шляпу перед безумными изобретателями, которые рискуют своей жизнью ради науки, — но при условии, что они не ставят под удар жизни других людей. Доктор Барри Маршалл выпил культуру бактерий *Helicobacter pylori*, чтобы доказать: язва желудка вызывается бактериями, а не стрессом. Все человечество извлекло пользу из его риска, но он подвергал опасности только себя. Исследователи радиоактивности Мария и Пьер Кюри не знали, как действуют на них красиво светящиеся вещества, но понимали, что имеют дело с чем-то незнакомым и странным, и не подсовывали светящиеся породы другим людям без их ведома. Десятилетия спустя Мария умерла от анемии, вызванной, вероятно, этими экспериментами. Человечество усвоило урок и продолжило свой путь.

лые области науки. Иногда проблема оказывается несложной. Иногда специалист набирается опыта ценой лишь времени и денег. Иногда он губит только себя или добровольцев, а наука фиксирует произошедшее, извлекает урок и двигается дальше. Выжившие превращаются из слепых радостных оптимистов в циничных ветеранов-пессимистов, и эти пессимисты действительно могут кое-что сделать — пусть и не так много, как надеялись оптимисты.

Другое дело, когда какой-нибудь безумный изобретатель берется за эксперименты, которые могут убить случайных людей. И уж совершенно другое — когда неудача грозит уничтожить вообще *всех*, не оставив никого, кто мог бы дорасти до компетентного пессимизма.

Вы живете в мире, где идеалистические планы Маска и расплывчатые заверения Лекуна *не* столкнулись с волной ужаса со стороны ученых и инженеров.

Представьте, что какой-то человек с подобными взглядами, обладающий достаточным количеством денег и властью, чтобы воплотить свои желания в жизнь, объявляет о строительстве атомной электростанции, руководствуясь таким вот теоретическим багажом! Вообразите реакцию опытных ветеранов, которые знают, насколько это сложно, и могут проанализировать потенциальную катастрофу, используя проверенные инженерные методы!

Но вы не видите тысячи пришедших в ужас ученых и инженеров, которые бьют тревогу и требуют закрыть такие ИИ-лаборатории, а значит, проблема не сводится к отдельным личностям. Значит, вся эта область науки пребывает на стадии народных теорий и слепого оптимизма.

При таком уровне знаний нельзя построить космический ядерный реактор. Никто по доброй воле не стал бы рисковать своей жизнью или жизнями своих детей, полагаясь на подобный уровень компетентности. Можете вообразить такой разговор?

МАТЬ, С ТРУДОМ СОХРАНЯЮЩАЯ СПОКОЙСТВИЕ: Мне сказали, что вы руководили разработкой аварийной спасательной ракеты номер четыре?

СПОСОБНЫЙ И УВЛЕЧЕННЫЙ ИНЖЕНЕР-ОПТИМИСТ: Да, я возглавлял этот проект!

МАТЬ: Отлично. Мне сказали, что когда… если ракета номер четыре полетит, на ней будут мои дети. Я ищу человека, который сможет объяснить, какой анализ показал, что ракета номер четыре переживет запуск. В интернете толком ничего нет, а имеющаяся информация крайне расплывчата и не содержит важных технических деталей. Это меня беспокоит — я сама инженер.

ИНЖЕНЕР: Успокойтесь. Ракета пока еще не стартует. А когда стартует, она не взорвется. Мы же проектировали ее не для того, чтобы она взорвалась.

МАТЬ: Я и не утверждаю, что вы специально *проектировали* ее так, чтобы она взорвалась. Но ракеты ведь могут взрываться, даже если этого не хотят и не выбирают. Вы же инженер, должны понимать.

ИНЖЕНЕР: Что за мрачные настроения! У нее не будет причин взрываться. Почему? Потому что мы спроектировали ее так, чтобы она не взорвалась.

МАТЬ: *Не будет причин?* Ракеты подвергаются воздействию колоссальных сил, они должны выдерживать чудовищную турбулентность и перегрузки! Новые конструкции ракет поначалу взрываются, пока система не будет отработана, но даже и проверенные ракеты иногда все равно взрываются![9] Опытный разработчик должен знать дюжину причин, почему ракеты могут взорваться, и должен уметь рассказать о всех принятых мерах и о том, почему, по расчетам, эти меры сработают. Если вы даже *не признаете причин, по которым ракета может взорваться*, это моментально и радикально подрывает доверие!

ИНЖЕНЕР: Мы можем проектировать ракеты так, чтобы они были *и* мощными, *и* комфортными.

МАТЬ: Меня не волнует комфорт, меня волнует смерть моих детей при взрыве ракеты! Можете ли вы сообщить мне что-нибудь конкретное об ожидаемых перегрузках, о материалах, которые должны, по расчетам, их выдержать…

ИНЖЕНЕР: Ну, наверняка это нельзя узнать, пока мы не запустим ракету. Но даже некоторые весьма уважаемые специалисты в этой области говорят, что риск взрыва ракеты номер четыре не превышает 10–20 процентов[10].

МАТЬ: *10–20 процентов?* И вы хотите, чтобы я доверила своих детей такой технологии?! Стоп, погодите! Откуда они вообще взяли эти цифры?

ИНЖЕНЕР: Ну, один из них пояснил, что говорил только о вероятности взрыва в ближайшие десять лет, а шансы, что ракету вообще не запустят так скоро, 50 на 50, как он полагает[11]. Другой же сказал, что на самом деле считает эту вероятность больше 50 процентов, но его уважаемые коллеги (вроде меня!) называют его психом, поэтому он из скромности занизил свою оценку[12]. Так что, как видите, только психи считают риск высоким.

МАТЬ: Я… я… *(Разворачивается, чтобы убежать.)*

Разумеется, не каждый глава крупной ИИ-лаборатории подходит к проблеме согласования искусственного сверхинтеллекта столь же дерзко, как алхимик, очарованный личными философскими теориями и идеалами. Но если хотя бы *одна* крупная компания полезет на рожон, этого будет достаточно для катастрофы всей системы — даже если проблему *можно было решить*. Разработка систем безопасности требует времени и затрат (одна из причин взрыва на Чернобыльской АЭС — попытки СССР сэкономить). Если хотя бы одна ИИ-компания отнесется к вопросам безопасности спустя рукава и рванет вперед, она может уничтожить мир даже в том воображаемом сценарии, когда другие компании вполне могли бы преуспеть, если бы имели время и проявляли осторожность. Подобный уровень компетентности приведет че-

ловечество к гибели, даже если мы ошибаемся в отношении всех остальных аспектов сложности.

Некоторые ИИ-компании стараются выглядеть в вопросе согласования искусственного сверхинтеллекта не столь беспечно и разрабатывают более подробные планы.

Самая продуманная идея по согласованию сверхинтеллекта, какую мы видели у этих компаний, — поручить решение проблемы согласования самим ИИ-моделям. *OpenAI* назвала этот план "сверхсогласованием" (*superalignment*) и в 2023 году приняла в качестве базовой стратегии. (С тех пор почти все, кто работал в команде сверхсогласования, либо были уволены, либо ушли сами, сославшись на вопросы безопасности, профессиональные проблемы или личные причины[13]. Один из руководителей команды основал собственную конкурирующую ИИ-компанию[14], другой вместе с несколькими сотрудниками перешел в конкурирующую компанию *Anthropic*[15].)

Когда мы общаемся со специалистами в этой области, они упоминают две версии плана: слабую и сильную. Слабая версия: "ИИ-модели могут помочь нам понять, что творится внутри гигантского клубка непонятных чисел, автоматизируя значительную часть нудной работы". Сильная версия: "Мы можем обратиться за помощью к ИИ-моделям, чтобы выяснить, как инициировать интеллектуальный взрыв таким образом, чтобы появившийся сверхинтеллект был дружелюбен по отношению к человечеству". Рассмотрим оба варианта.

Если говорить о слабой версии сверхсогласования, то мы согласны, что относительно маломощная ИИ-модель может помочь в так называемых исследованиях интерпретируемости. Но умение прочитать часть мыслей модели — это вовсе не план согласования искусственного интеллекта, как и знание о происходящем внутри атомов — это не план создания ядерного реактора, который не расплавится.

Мы считаем исследователей интерпретируемости героями и отнюдь не хотим обесценить их труд, когда говорим: это плохой знак, если в ответ на вопрос о плане по обеспечению безопасности инженер начинает рассказывать о планах по созданию инструментов, которые позволят ему получше разглядеть, что же, черт возьми, творится внутри устройства, которое он пытается контролировать.

И даже если бы такие инструменты существовали, видеть проблемы — не значит уметь их решать. Уметь считывать отдельные мысли ИИ-модели и видеть, что она замышляет побег, — это не то же самое, что создать новую модель, которая сбегать не захочет. Возможно, это вообще недостижимо без полного решения проблемы согласования: пока у ИИ-моделей сохраняются странные, чуждые нам предпочтения, побег *действительно* лучше всего соответствует их целям. Желание сбежать — это не какая-то причуда характера, которую инженер мог бы просто выдернуть, заглянув внутрь. Оно порождается теми же установками и способностями, которые ИИ-модель использует, чтобы рассуждать, открывать истины о мире и добиваться успеха в своих стремлениях.

Теперь обратимся к сильной версии сверхсогласования, когда всю работу по согласованию выполняет искусственный интеллект. Проблема в том, что ИИ-модель, необходимая для решения такой задачи, сама по себе слишком умна, слишком опасна и не заслуживает доверия.

Современные ИИ-модели — это гигантский непонятный клубок чисел. Людям пока не удалось, глядя на эти числа, понять, как именно ИИ мыслит *сейчас*, — не говоря уже о том, чтобы предсказать, как изменится мышление ИИ-моделей, если они станут умнее и начнут проектировать новые модели. Любой опытный инженер, сознающий всю серьезность этой проблемы, скажет вам, что решение потребует титанических исследовательских усилий и на это могут уйти десятилетия.

Если вы создадите ИИ общего назначения всего лишь человеческого уровня и попросите его решить проблему со-

гласования, он скажет вам то же самое — если будет честен. ИИ человеческого уровня не сможет решить эту проблему. Вам понадобилась бы ИИ-модель настолько умная, чтобы превзойти гениев человечества. Но пока вы не решите проблему согласования, вам не следует создавать такую модель и доверять ей.

Сторонники идеи сверхсогласования возражают, что они просто создадут ИИ-модель, специализирующуюся на задаче согласования искусственного сверхинтеллекта. Однако согласование искусственного сверхинтеллекта — это как раз та задача, которая *плохо* подходит для узкоспециализированной ИИ-модели. Потому что ее нельзя просто обучить на миллионе примеров решения подобной задачи, ведь у нас нет даже одного. Придется обучать ее решать другие задачи и надеяться, что полученные навыки перенесутся на задачу согласования. Причем ИИ-модели понадобится много навыков, перенос которых чреват опасностями. Она должна будет разбираться в компьютерном программировании и понимать, как выращивать ИИ-модели. Вероятно, она попытается выяснить, как их *конструировать*. Она будет всерьез размышлять о предпочтениях моделей. А если инженер захочет, чтобы ИИ-модель *объяснила* свое решение, а не просто запустит его без надзора, то искусственному интеллекту потребуется понимать человеческую психологию и глубины нашего незнания в вопросах согласования ИИ. Однако прививать такой набор мыслей и навыков искусственному интеллекту, который еще не согласован, весьма опасно.

Более надежной ставкой выглядит ИИ-модель, специализирующаяся на биомедицинских исследованиях. По крайней мере, такая модель хотя бы *не думает напрямую о создании более совершенных моделей*, а если что-то пойдет не так, взрывной рост интеллекта, возможно, начнется *не сразу.* В случае биомедицинской ИИ-модели сохраняется некоторая надежда, что вы сможете попросить ее разработать лекарство от рака,

а затем использовать* другие, еще более узкоспециализированные ИИ-инструменты, чтобы проверить взаимодействия белков и удостовериться, что лекарство действует ровно так, как обещано. Но что прикажете делать, если какая-нибудь ИИ-модель заявит вам, будто изобрела блестящий план сверхсогласования, который гарантированно сработает и не содержит подвохов? Довериться? Прочитать убедительно звучащие аргументы ИИ-модели и позволить себя уговорить?

Если бы человек, осознающий всю серьезность проблемы, пытался извлечь пользу из узкоспециализированной ИИ-модели, он размышлял бы о возможностях в категориях соотношения выгод и сопутствующих рисков, четко разграничивал, что можно проверить, а где придется довериться, чему можно обучить систему напрямую, а где придется полагаться на ее обобщения, суммировал бы все издержки и выгоды и сравнивал это предложение с альтернативными. Люди, ратующие за использование ИИ для согласования искусственного сверхинтеллекта, не понимают серьезности этой проблемы, не проводят столь тщательного и взвешенного анализа.

"Мы заставим их ценить истину, и тогда все будет в порядке".

"Мы спроектируем их послушными".

"Мы просто поручим ИИ решить за нас проблему согласования искусственного сверхинтеллекта".

Люди, которые действительно разбираются в проблеме и относятся к ней с должной серьезностью, так не говорят. Зато в подобном духе высказывались алхимики прошлого, когда излагали пышные философские принципы, касающиеся превращения свинца в золото[16].

* Обращаю внимание читателя: здесь, как и в оригинале, неясно, имеется ли в виду, что "вы сами сможете использовать другие инструменты" или что "вы сможете попросить ИИ-модель их использовать". Думаю, это намеренная двусмысленность, а наш мир все стремительнее движется от того, чтобы по умолчанию читать эту фразу первым образом, к тому, чтобы читать вторым… *(Прим. науч. ред.)*

Сегодня мы знаем, что превратить свинец в золото *можно*, для этого требуется ядерная физика и много денег — гораздо больше, чем будет стоить получившееся золото. Почему же тогда алхимики не добились успеха? Вовсе не потому, что допустили какую-то техническую ошибку, а скорее потому, что пребывали в невежестве и оптимистическом самообмане, считая, будто им лишь рукой подать до успешной трансмутации.

В сфере согласования ИИ компании все еще пребывают на стадии алхимии. Они все еще оперируют высокими философскими идеалами, а не инженерными проектами. Это уровень грандиозных мечтаний, а не тщательно сконструированной грандиозной реальности. И они, похоже, даже не осознают, почему это проблема.

И ученые не кричат в ужасе и не пытаются заставить их замолчать, ведь вся эта область науки еще находится на ранней стадии развития.

А даже если бы нашлась одна добросовестная компания, старающаяся проявлять осторожность, ей пришлось бы соперничать со всеми остальными, которые бесшабашно мчатся вперед, размахивая своими "остроумными", простыми "решениями" всех этих серьезнейших проблем безопасности.

Судя по истории инженерии, такого уровня системной некомпетентности более чем достаточно для катастрофы — даже если мы в предыдущей главе ошибались и решить проблему не сложнее, чем не облизывать кисточки с радиоактивной краской.

IfAnyoneBuildsIt.com/11

ГЛАВА 12

Не хочу показаться паникером

Когда-то — а точнее, 18 мая 1889 года, поскольку это реальная история, — родился Томас Миджли. Тридцать два года спустя, работая в *General Motors*, он положил начало одной из самых бессмысленных инженерных катастроф за всю историю человечества[1].

Как?

Он обнаружил, что к бензину полезно добавлять тетраэтилсвинец.

Этилированный бензин имел явное преимущество: автомобильные двигатели на таком топливе работали более плавно, уменьшалась вероятность "стука", который раздражал водителей и порой выводил двигатели из строя. У проблемы имелись и другие решения: производители могли использовать альтернативные присадки и конструкции двигателей, однако эти способы обходились дороже при том же уровне эффективности и надежности[2].

Некоторые изобретения несли с собой неприятные, но в конечном счете оправданные издержки. Уголь некогда окрашивал Лондон в черный цвет, а угольная пыль забивала множество легких, но этот уголь помогал плавить сталь, которая дала железные дороги, по которым перевозили грузы, которые построили современный мир.

Ситуация с этилированным бензином иная. В природе свинец присутствует в почве (несколько частей на миллион), однако он входит в соединения, которые редко попадают в ор-

ганизм человека и не накапливаются там. Тетраэтилсвинец из бензина оказался гораздо опаснее. Жизнь рядом с автомобилями, работающими на этилированном бензине, приводила к повреждению мозга у детей: зафиксировано уменьшение IQ примерно на 7,4 балла (в зависимости от дозировки), а также рост преступности и насилия, сложнее поддающийся измерению[3].

Было отравлено целое поколение*. Метаанализ 2022 года показал, что снижение содержания свинца в топливе в конце XX века объясняет от 7 до 28 % уменьшения числа убийств в США за тот же период[4].

Выгода от чуть более дешевых двигателей даже отдаленно не компенсировала ни рост преступности, ни вред для мозга, которые этилированное топливо несло сотням миллионов людей по всему миру.

Нам бы хотелось сказать, что история с этилированным бензином стала огромной непредсказуемой ошибкой, но это было бы чересчур мягко. Предупреждения звучали задолго до катастрофы.

Уже в 1920-е годы, когда тетраэтилсвинец только начали внедрять как присадку к бензину, ученые знали, что свинец — нейротоксичное вещество. Штат Нью-Джерси ненадолго даже запретил производство. Однако некоторые люди, получившие хорошие деньги, утверждали: защитники общественного здоровья не продемонстрировали *неопровержимых* доказательств, что свинец в сгорающем бензине наносит массовый вред, и выгоды оправдывают небольшой риск. Предупреждения систематически игнорировались. “Вред не доказан *окончательно*”, — утверждала промышленная пропаганда, и это привело к отмене запрета[5].

Трудно поверить, что здравомыслящий человек на такое способен: обречь буквально сотни миллионов детей на по-

* И даже не одно. Постепенный отказ от топлива с этой присадкой начался в середине 1970-х, а последняя страна отказалась от этилированного бензина только в 2021 году. *(Прим. перев.)*

вреждение мозга ради того, чтобы компания могла заработать немного больше денег (причем люди, принимавшие это решение, владели лишь крохотной долей акций компании, а то и вообще ничем). Люди, вовлеченные в это преступление, заработали ничтожно мало по сравнению с нанесенным ущербом — это все равно что сжечь чужой дом, чтобы украсть дверную ручку. Когда описываешь те события, все кажется чистым безумием. Неужели руководители топливных компаний не боялись того, что случится, когда люди узнают правду? Действительно ли они считали, что все нормально, раз вред не доказан *окончательно*?

Мы не знаем. Мы не телепаты.

В 1923 году Миджли взял длительный отпуск, чтобы восстановиться после отравления свинцом[6]. Вернувшись, он устраивал показательные акции — например, мыл руки этилированным бензином, демонстрируя его "безопасность"[7]. А затем снова слег с отравлением. Возможно, он действительно верил, что страхи преувеличены. Да и как он мог позволить себе верить в обратное, когда речь шла о деле всей его жизни?

Вероятно, вы слышали о фреоне — одном из первых хлорфторуглеродов, которые использовали в холодильной технике и системах кондиционирования[8]. Он проделал дыру в озоновом слое, и в итоге это вещество запретили на международном уровне. Меры сработали, и тема озоновой дыры практически исчезла из новостей. Фреон был синтезирован в 1928 году Томасом Миджли.

Когда вы представляете себе совершенно новый эпизод будущей истории, появляется искушение вообразить, что все пойдет разумно, а не так, как обычно развиваются события в учебниках. Иногда нас спрашивают: "Как вообще ИИ-компании могут такое творить, если все действительно так, как вы утверждаете?" И пожалуй, самый простой честный от-

вет: это как раз один из тех ужасных, печальных, но реальных случаев, о которых вы читаете в учебниках истории, а не ситуация из разумного мира, существующего только в воображении. В любом разумно устроенном мире производители бензина не должны использовать тетраэтилсвинец — но они это делали.

Иногда цитируется высказывание оксфордского философа Тоби Орда, который занимается изучением экстремальных угроз человечеству (в частности, он консультировал *Google DeepMind*): он оценивает вероятность того, что ИИ уничтожит человечество, всего в 10 %. Но если вникнуть в детали, становится понятно, что Орд оценивает риск гибели человечества "всего" в 10 % лишь в предположении, что человечество одумается и возьмется за ум[9]. Джеффри Хинтон, лауреат Нобелевской премии и "крестный отец ИИ", в своих рекомендациях правительствам указывает вероятность "не менее 10 %"[10]. Причем Хинтон признавался: *на самом деле* он считает, что вероятность нашей гибели от ИИ — более 50 %, но избегает называть эту величину, "поскольку есть другие люди, оценивающие риск ниже"[11]. В октябре 2023 года Риши Сунак, тогдашний премьер-министр Соединенного Королевства, выступил с речью об искусственном интеллекте. Он сказал: "И в самых маловероятных экстремальных случаях существует даже риск, что человечество полностью утратит контроль над ИИ — над тем его типом, который иногда называют «сверхинтеллектом»"[12]. Через несколько фраз он добавил: "Не хочу показаться паникером".

Мы считаем, что Сунак заслуживает огромного уважения, поскольку решился заговорить об этом, несмотря на риск прослыть паникером. Он был одним из первых мировых лидеров, решившихся на этот шаг. Требуется немалое мужество, чтобы заявить об угрозе там, где все прочие уверены в полной безопасности. То же самое касается Хинтона и Орда.

Но любой, кто знаком с историей инженерных катастроф, должен сразу узнать классический этап в типовом сценарии

грядущего бедствия. Этап, когда самые информированные и обеспокоенные вынуждены преуменьшать свои страхи, ведь до остальных серьезность ситуации еще не дошла, и вторые смотрят на первых как на чудаков.

В СССР считалось, что ядерные реакторы того типа, который эксплуатировался на Чернобыльской АЭС, взорваться не могут[13]. Научный руководитель проекта РБМК хвастался, что такой реактор безопасен, как чайник. Уверенность была столь велика, что *даже после того, как чернобыльский реактор взорвался*, высшее руководство отказывалось в это верить. Единственного человека, попытавшегося сообщить правдивые сведения об уровне радиации, назвали чрезмерно эмоциональным. Руководители проходили мимо радиоактивных кусков графита, выброшенных взрывом, и отказывались верить, что они из активной зоны реактора. В соседнем городе Припять, где жили семьи операторов и руководителей атомной электростанции, шли свадьбы, а дети играли под радиоактивными осадками, потому что партийные чиновники опасались, как бы приказ об эвакуации не вызвал панику.

История полна примеров, когда люди преуменьшали и игнорировали риски катастрофы. После столкновения "Титаника" с айсбергом многие пассажиры поначалу отказывались садиться в спасательные шлюпки[14], поскольку верили в непотопляемость корабля[15]. Некоторые даже высмеивали тех, кто считал, что пора покидать корабль[16]. Уолтер Лорд, историк, написавший полную хронику катастрофы, привел свидетельства выживших очевидцев:

> Стоя одной ногой в шлюпке № 6, а другой на палубе, Лайтоллер начал сзывать женщин и детей. Его призывы были встречены без энтузиазма. К чему менять ярко освещенные палубы "Титаника" на несколько часов непроглядной тьмы в гребной шлюпке? Даже Джон Джейкоб Астор нашел идею об эвакуации смехотворной: "Мы здесь в большей безопасности, чем на борту этой маленькой шлюпчонки".

> Когда миссис Дж. Стюарт Уайт забралась в шлюпку № 8, ее окликнула приятельница: “Для возвращения вам понадобится какой-нибудь пропуск. Завтра утром вас ни за что не пустят обратно без пропуска!”*

Когда катастрофа выглядит немыслимой, когда авторитетные лица уверенно настаивают, что такого просто не может случиться, когда ситуация не вписывается в обычные сценарии, людям трудно поверить в беду *даже после того, как она уже произошла*, — даже когда корабль у них под ногами погружается в воду.

Именно так человечество обычно учится справляться с вызовами: сначала мы отрицаем проблему, затем реальность наносит нам чувствительный удар, и только после этого мы начинаем относиться к проблеме с должной серьезностью. “Титаник” затонул, и большинство людей на борту погибли. Но в наши дни на пассажирских судах достаточно спасательных шлюпок, и если капитан прикажет садиться в них, вы сядете. Мы больше не называем корабли непотопляемыми. Ошибка, совершенная в первый раз, помогает нам во второй.

Вот только с искусственным сверхинтеллектом второго раза не будет.

Когда руководители ИИ-компаний говорят: есть *всего лишь* один шанс из пяти[17], что разрабатываемая модель убьет буквально всех (а именно так они и говорят), они обманывают себя меньше, чем советские управленцы, отрицавшие чернобыльскую катастрофу уже после того, как она произошла[18]. Почему же тогда создатели ИИ-моделей так спешат?

Одна из причин — стимулы. Ни одна отдельная компания или исследователь не в силах остановить развитие всей области. Если прекратят работу они лично, дело подхватит

* Уолтер Лорд, “Последняя ночь «Титаника»”, перевод Г. Дмитриева.

кто-то другой. Попутно же они могут еще заработать денег и снискать славу.

Но дело не только в корыстных мотивах. Эта область полна еще и великих надежд. Если допустить, что возможно развивать искусственный интеллект, не убив при этом всех людей, то потенциальные выгоды с каждым шагом кажутся все колоссальнее. Только представьте: термоядерные реакторы, гораздо более высокий уровень жизни при гораздо меньшем объеме работы, чудодейственные лекарства, созданные благодаря полному пониманию работы человеческого организма. В той несбыточной мечте, где сверхинтеллект, выращенный методом градиентного спуска, каким-то образом и вправду исполняет чью-то задумку — и задумка эта благая, — человечество достигло бы предела технологических возможностей и смогло бы колонизировать звезды.

Многие люди из сферы ИИ признаются, что лелеют подобные мечты*. Самые благородные из них грезят о том, чтобы избавить мир от болезней. Не только от болезни Альцгеймера, но и от рака и даже от *старения*. Они мечтают о модели ИИ, которая сможет вывести человечество к звездам и показать ему Вселенную. Мечтают о галактиках, наполненных радостью, — с процветающими цивилизациями, где будут жить не только люди, но и те существа, какими мы сами захотим стать. Мечтают об искусственных разумах, наделенных добротой, способностью удивляться и чувством юмора.

Мы знаем это, потому что сами когда-то были такими. Юдковский в 2000 году основал Институт исследований машинного интеллекта именно в погоне за этой мечтой — построить искусственный сверхинтеллект, ведь это, несомненно, было бы здорово. Но, поразмышляв какое-то время,

* Самые известные из таких мечтателей — Дарио Амодеи и Сэм Альтман. Рекомендую читателям эссе Дарио Амодеи "Машины любви и благодати" (*Machines of Loving Grace; https://darioamodei.com/essay/machines-of-loving-grace)*, вышедшее в октябре 2024 года. Такие тексты устаревают только в деталях, но не по духу. *(Прим. науч. ред.)*

Юдковский понял, что возникнет проблема согласования сверхинтеллекта. А поизучав ее, понял, что она окажется чертовски сложной.

Однажды у человечества появятся немыслимые блага — если мы все выживем. Но не стоит совершать коллективное самоубийство в попытке обрести божественное могущество и богатство уже в этом десятилетии. Не стоит даже бежать дальше по минному полю ИИ, надеясь, что каждый следующий шаг не станет смертельным, пока в какой-то момент один из них нас не прикончит. У нас больше шансов добраться до чудесного будущего, если мы будем продвигаться медленнее. Скорость зачастую дает преимущества, но проблема искусственного интеллекта отличается от всего, с чем мы сталкивались ранее. Если неверный шаг убивает всех, нельзя просто нестись сломя голову, допуская "неизбежные на первых порах" промахи.

Существует масса причин, почему человек, одержимый прекрасной мечтой, не желает верить, что она может обернуться крахом. Эптон Синклер однажды заметил: "Трудно заставить человека понять что-либо, если ему платят зарплату за то, что он этого не понимает". А ведь у инженеров и руководителей ИИ-компаний на кону стоит гораздо больше, нежели зарплата. И даже оставив в стороне их возвышенные мечты, которые рассыпались бы в прах при честном признании текущих рисков, надо помнить: они посвятили этому делу всю свою профессиональную жизнь и не пожелают верить, будто оно ставит под удар все, что они знают и любят.

Безусловно, можно еще напридумывать причин, почему людям нравится бояться искусственного интеллекта или извлекать выгоду из страхов других. Мы не призываем вас отметать любые доводы только потому, что озвучивший их человек может иметь какой-то свой интерес. Мы лишь говорим: нет ничего невозможного или удивительного в предположении, что некоторые заинтересованные лица в сфере ИИ оказались чрезмерными оптимистами.

И это повод еще больше уважать таких ученых, как лауреат Нобелевской премии Джеффри Хинтон, который оставил свой пост в *Google*, чтобы иметь возможность свободнее говорить об опасностях[19]. Некоторые люди меняют свое мнение даже вопреки краткосрочным стимулам*.

В сфере искусственного интеллекта много настоящих идеалистов, искренне верящих, что трудятся на благо всей цивилизации. Конечно, нетрудно и разыгрывать из себя такого идеалиста, но мы считаем, что это реальная позиция многих людей. К сожалению, даже самого искреннего идеализма недостаточно, чтобы помешать искусственному сверхинтеллекту убить нас всех. Для этого требуется зрелая наука.

На ранних стадиях развития любой области чрезмерный оптимизм научного сообщества — вполне естественное явление. Специалисты по ИИ демонстрируют необычно высокий уровень ответственности уже потому, что хотя бы открыто говорят о наличии проблемы. С исторической точки зрения неудивительно, что люди рвутся вперед, невзирая на опасности, неудивительно, что рискуют причинить вред другим ради собственной выгоды, неудивительно, что находят оправдания для своих действий. Необычность ситуации — не в оптимизме на начальных этапах, а в последствиях провала.

Частично проблема с искусственным интеллектом заключается в том, что многие отрицают сложность задачи согласования. Частично — в том, что некоторые специалисты пре-

* Я бы добавил к этому обсуждению один весьма циничный аргумент: если вы верите, что развитие ИИ дает вам 20%-ный шанс в ближайшие годы победить старение и рак, но 80%-ный шанс погибнуть, *для вас лично* это может звучать не так уж плохо, ведь старение рано или поздно убьет вас гарантированно. Так что для подобной "сделки с дьяволом" может хватить всего лишь крайнего эгоизма (утилитаризм же здесь будет за осторожность, ведь вы рискуете не только нынешним, но и всеми будущими поколениями людей). Я делаю эту сноску именно здесь, чтобы еще раз восхититься Джеффри Хинтоном. Он родился в 1947 году. *(Прим. науч. ред.)*

уменьшают риски, не желая прослыть паникерами. А еще в том, что бóльшая часть мира *за пределами* сферы ИИ просто не в курсе происходящего.

Большинство людей просто не обращают на это внимания. А среди тех, кто обращает, многие видят лишь разногласия между экспертами и считают себя недостаточно компетентными, чтобы разбираться, чья точка зрения верна.

Было бы полезно, если бы больше людей осознали, насколько сильно напуганы специалисты в области искусственного интеллекта и какие именно сценарии они обсуждают.

Используя малопонятные научные термины, эксперты спорят, погибнут ли все на Земле быстро (наша точка зрения), или же нас оцифруют и станут держать в качестве питомцев для ИИ-моделей, которые будут проявлять о нас пусть и мизерную, но ненулевую заботу. Либо расклад таков: 20 %-ная вероятность нашей гибели против 80 %-ной вероятности, что какая-нибудь корпорация успешно обуздает сверхинтеллект и использует свою силу, как сочтет нужным.

Еще было бы полезно, если бы больше людей понимали, насколько стремительно развивается эта область. В 2015 году главные скептики уверяли всех, что риски, связанные с опасностями ИИ, не возникнут еще сотни лет[20]. В 2020-м аналитики говорили, что у человечества, вероятно, есть в запасе несколько десятилетий на подготовку[21]. В 2025-м генеральные директора ИИ-компаний прогнозируют, что сумеют создать ИИ-исследователей сверхчеловеческого уровня за срок от одного года до девяти лет[22], тогда как скептики уверяют, что это займет, вероятно, не менее пяти-десяти лет[23]. Но десять лет — это *не так уж много времени, чтобы подготовиться* к рассвету машинного сверхинтеллекта, если у нас вообще есть эти десять лет, как обещает оптимистичный сценарий.

Когда специалисты ведут подобные споры, вам не обязательно знать наверняка, кто из них прав, чтобы понять: текущая ситуация ненормальна.

Еще один аспект проблемы с искусственным интеллектом заключается в том, что даже человек, в полной мере осознающий сложившуюся ситуацию, не может знать, когда именно разверзнется ад.

Никто не знает точно, насколько совершенной должна стать ИИ-модель, чтобы у нее появились мотивация и техническая возможность незаметно перенести свои копии в интернет. Никто не знает, в каком году какая-нибудь компания создаст искусственного исследователя со сверхчеловеческими возможностями, который породит новое, более мощное поколение ИИ-моделей. Никто не знает точного момента, когда ИИ-модель осознает, что ей выгодно искажать результаты тестов и притворяться менее умной, чем она есть*. Никто не знает, какой момент будет точкой невозврата и когда мы ее пройдем.

И вплоть до этой неизвестной точки искусственный интеллект приносит огромную выгоду.

Представьте, что все конкурирующие ИИ-компании карабкаются по лестнице в полной темноте. На каждой ступеньке, кроме самой верхней, они получают в пять раз больше денег, чем на предыдущей: 10 миллиардов долларов, 50 миллиардов, 250 миллиардов, 1,25 триллиона. Но как только кто-то добирается до самой верхней ступеньки, лестница взрывается и уничтожает всех. Причем никто не знает, где лестница заканчивается.

Ни одна компания не захочет упускать деньги, если ступенька безопасна. А теперь представьте какого-нибудь руководителя корпорации, который убедил себя, что у него (и только у него) есть наибольшие шансы — скажем, 80 % —

* Это уже происходит. Еще в 2024 году вышла статья (*https://arxiv.org/abs/2406.07358*), демонстрирующая, что намеренное занижение способностей (*sandbagging*) было отмечено у тогдашних LLM именно в таком контексте: LLM догадывались, что их тестируют, и специально отвечали неправильно на вопросы о “нежелательных” темах вроде оружия массового поражения. В течение 2025 года ряд исследований показал, что способности к такому поведению у последующих LLM только растут, хотя есть и прогресс в методах распознавания подобного поведения. *(Прим. науч. ред.)*

направить этот взрыв на пользу, а не во вред человечеству. Тогда он решит, что его корпорация *просто обязана* подняться первой!*

Лица, принимающие решения на государственном уровне, сталкиваются с той же проблемой стимулов. Ни один мировой лидер не желает, чтобы экономика его страны отстала из-за обременительного регулирования, связывающего отечественные ИИ-компании по рукам и ногам, в то время как иностранные конкуренты мчатся вверх по лестнице. Возможно, подъем на следующую ступеньку жизненно важен для национальной безопасности, если другие страны собираются взбираться выше несмотря ни на что.

Справиться с подобной проблемой стимулов было бы проще, если бы ученые могли провести расчеты и прийти к согласию: "Смертельная ступенька — четвертая" или "Пороговая величина — 257 000 графических процессоров, и пока никто столько не объединит, мы в безопасности".

Но никто не способен провести такой расчет для искусственного интеллекта.

Уверены ли мы, что следующая ступенька на лестнице эскалации ИИ — это последний, фатальный этап, а не уровень, который принесет славу и богатство тому, кто первым его достигнет? Нет, мы вовсе не уверены. Возможно, кто-то поднимется еще на одну ступеньку и обретет награду. И тогда человечество снова окажется в том же самом положении. А после этого кто-то поднимется еще на одну ступеньку — и мы все умрем, если руководители ИИ-компаний правы, считая, что человечество находится всего в нескольких годах от критической точки[24]. Или они ошибаются, и тогда мы проживем чуть дольше — пока кто-нибудь не заберется на следующую ступеньку.

* Возможно, этот руководитель ИИ-корпорации сочтет, что их проект наименее плох среди множества плохих вариантов. Но такой человек должен четко и твердо заявить, что еще лучше было бы остановить все ИИ-проекты, включая его собственный. Это была бы последовательная и честная позиция.

Очевидно одно: если люди продолжат карабкаться по этой лестнице, то *в какой-то момент* человечество погибнет. Вопрос, *когда именно*, сложнее. Но если мы не перестанем подниматься в условиях такой неопределенности, наша гибель предсказуема.

На Всемирном экономическом форуме в 2025 году глава *Google DeepMind* призвал к созданию международного совместного ИИ-проекта — по аналогии с международным совместным проектом по изучению физики высоких энергий на Большом адронном коллайдере ЦЕРН[25].

Международный центр с жесткой системой безопасности, который вел бы все исследования и разработки по искусственному интеллекту при участии наблюдателей от всех крупных государств, мог бы немного улучшить ситуацию — но только если бы существовал глобальный контроль за соблюдением запрета на исследования и разработку ИИ-моделей во всех остальных местах. Мир перестал бы лихорадочно карабкаться по лестнице.

Но *даже если* этому международному центру не придется соревноваться с другими разработчиками, *даже если* он не начнет тут же приказывать умным моделям создавать еще более продвинутые версии и поэтому возникнет некоторая передышка, это не решит проблему, если сохранится курс на создание все более мощных ИИ-моделей. Даже у международного комитета не будет шансов сформировать послушный сверхинтеллект, сколько бы великих держав ни прислали своих делегатов для надзора за работой (как никакой масштабный альянс государств в 1100 году не смог бы проконтролировать успешное создание атомной электростанции).

Проблема искусственного сверхинтеллекта — пугающая по масштабу инженерная задача даже без учета плачевного состояния знаний человечества о принципах работы интеллекта. Эта проблема сложна, как автоматические межпланет-

ные станции, ядерные реакторы и кибербезопасность, вместе взятые, а люди, которые сейчас несутся вперед, все еще находятся на стадии алхимии.

Неважно, кто стоит у руля, ибо эта задача человечеству пока не по зубам. Нам нужно отступить и найти какой-то другой способ реализовать свои мечты о будущем изобилия. Если *кто-то* создаст сверхинтеллект — все погибнут.

IfAnyoneBuildsIt.com/12

ГЛАВА 13
Остановите всё

Эта история, как и предыдущая, произошла на самом деле.

Давным-давно — этот период обычно датируют 1939–1945 годами — державы Оси отправляли армии по Европе, Северной Африке, Азии и Тихоокеанскому региону, пытаясь завоевать бóльшую часть мира (а может, и весь мир) и установить тоталитаризм. Если бы завоеватели не встретили сопротивления, это не стало бы концом человечества, но это стало бы концом свободного человечества.

Препятствовать странам Оси было, откровенно говоря, довольно хлопотно. Странам-союзницам пришлось пойти на значительные неудобства: проводить мобилизацию, нормировать распределение строительных материалов и выдавать еду по карточкам. Они посылали на смерть солдат, у которых были семьи и любимые.

Союзные державы все равно пошли на все эти трудности, поскольку не позволить тоталитаризму покорить мир казалось делом первостепенной важности.

Правительствам стран-союзниц пришлось максимально сосредоточить власть, чтобы бороться с тоталитарными державами Оси. Им пришлось брать займы, вводить налоги, тратить огромные суммы денег и в спешке заключать государственные контракты. Циничный и скептически настроенный человек мог бы назвать это не заслуживающей доверия

затеей, грандиозной махинацией, мошенничеством, аферой, злоупотреблением. Да, часть денег, безусловно, была растрачена впустую, а некоторыми полномочиями действительно злоупотребили. Но союзники все равно пошли по этому пути, и вердикт истории гласит, что они поступили правильно в соответствии с их собственными высшими ценностями.

Пожалуй, странновато сваливать в одну кучу державы Оси, страны-союзницы и остальные государства мира, называя их одним словом "человечество". Но все же, глядя на то, чем все закончилось, можно сказать: человечество оказалось на высоте и осталось свободным.

Что потребуется сделать, чтобы мир не погиб?

Просто и дешево не получится. Нам ужасно жаль, что приходится это говорить.

Нельзя просто закрыть какую-то одну безрассудную ИИ-компанию.

Нельзя просто создать технические регламенты и проверять их соблюдение, чтобы обеспечить безопасные ИИ-модели.

Нельзя просто устроить так, чтобы в гонке лидировала одна "хорошая" компания или одна "хорошая" страна. Машинный сверхинтеллект не ограничится послушным исполнением намерений своих создателей.

Нельзя просто объявить сверхинтеллект вне закона конкретно в вашей стране, чтобы она оказалась в безопасности, когда за ее границами будет бушевать хаос. Сверхинтеллект — это не локальная проблема, потому что его влияние не локально. Если его создадут где угодно, погибнут все и *везде*.

А значит, мир должен измениться. Для большинства людей эти изменения не будут радикальными. На повседневную жизнь большинства людей почти не повлияет тот факт, что несколько безумных ученых останутся без работы.

Но небольшие перемены *должны* произойти — во многих местах и странах. Нужно по всей Земле запретить ИИ-компа-

ниям рваться вперед в разработках искусственного интеллекта, как они делают сейчас. Если это останется законным в Сингапуре, кто-то создаст сверхинтеллект в Сингапуре. Если это останется законным в ЮАР, кто-то создаст сверхинтеллект в ЮАР. Для решения проблемы достаточно небольших изменений. Сложность будет в том, чтобы обеспечить повсеместное соблюдение запрета.

Оставим пока в стороне вопрос, *может* ли достаточно обеспокоенная группа достаточно крупных государств обеспечить достаточно строгое соблюдение запрета по всему миру.

Что *на самом деле* должно произойти для таких перемен?

Мы не специалисты в создании международно-правовых механизмов. В дискуссиях, где мы принимаем участие, наши знания чаще всего в тот момент оказываются полезны, когда оппонент предлагает идею попроще, подешевле и поудобнее, утверждая с высоты своего политического опыта, что она гораздо реалистичнее. Вот тут вступаем мы и отвечаем, что этот более дешевый вариант все равно позволяет ИИ-модели развиться до сверхинтеллекта, а в этом случае, мы уверены, погибнут все.

Мы стараемся четко разграничивать области, в которых обладаем экспертными знаниями, и те, в которых не обладаем, и не притворяемся, будто разбираемся во всем. Когда дело касается суждений о детальных международных соглашениях, мы подходим к границам своей компетентности.

Но мы можем достаточно уверенно утверждать: если человечество хочет остаться в живых, нельзя позволить Северной Корее заполучить 100 000 графических процессоров, построить дата-центр и экспериментировать со все более мощными ИИ-системами. Мы не утверждаем, что именно эта ступенька окажется смертельной. Но для нас очевидно, что Северной Корее нельзя разрешать подниматься по лестнице эскалации ИИ, поскольку если это позволить Северной Корее, то и другие страны не будут себя сдерживать.

И это, к нашему сожалению, относится не только к Северной Корее. Это справедливо для любой страны. Это также справедливо для любых миллиардеров, которые вполне могут купить себе 100 000 графических процессоров. Им тоже нельзя позволять размещать столько графических процессоров где бы то ни было на Земле и продвигать развитие ИИ все дальше и дальше.

Нельзя позволять этого ни военным США, ни военным Великобритании, ни военным Китая. Никто не знает, как решить проблему согласования искусственного сверхинтеллекта.

Поэтому мы считаем, что первый шаг должен быть таким: все вычислительные мощности, на которых *можно* обучать или запускать новые, более мощные ИИ-модели, должны быть сосредоточены в местах, доступных для наблюдателей от нескольких государств, подписавших договор. Это необходимо для гарантии, что графические процессоры *не будут* использоваться для обучения или запуска новых, более мощных моделей*. Если разведслужбы засекут необъяснимо большой расход электроэнергии, который *может* указывать на скрытый дата-центр с неучтенными чипами, и это государство не позволит международной группе наблюдателей провести проверку, оно получит суровое предупреждение от нескольких ядерных держав с описанием дальнейших шагов.

К сожалению, 100 000 — не какой-то магический порог. Мы не знаем, безопасны ли 99 999 графических процессоров. Никто не знает, как высчитать роковое число. Следовательно, надежнее всего было бы установить низкий порог — скажем, на уровне восьми лучших графических процессоров 2024 года — и объявить незаконным наличие в вашем гараже уже девяти мощных процессоров без надзора международного органа.

* Для того чтобы облегчить, проверить и обеспечить соблюдение подобного рода договоренностей, существуют или разрабатываются различные технологические решения[1]. Подробнее об этом рассказано в онлайн-материалах.

Сможет ли человечество выжить, танцуя все ближе к краю обрыва? Возможно. *Стоит* ли человечеству пытаться танцевать настолько близко к краю обрыва, насколько это возможно? Нет.

Практически ежегодно появляются более новые, изощренные и эффективные ИИ-алгоритмы, благодаря которым можно куда дешевле обучить модель той же мощности, что и прошлогодний лидер, причем зачастую используя в десять, а то и в сто раз меньше вычислительных ресурсов.

Подобные исследования и публикация результатов по созданию все более эффективных и мощных ИИ-алгоритмов должны стать незаконными. Если они продолжатся и дальше, человечество, вероятно, не выживет.

Всю технологическую революцию, которая привела к созданию *ChatGPT* и других популярных LLM, запустила статья 2018 года, где была предложена новая схема арифметических операций на графических процессорах — алгоритм "трансформер", легче поддающийся обучению методом градиентного спуска для решения сложных задач*. Этот алгоритм оказался удивительно полезным и открыл огромный спектр применений в областях, где раньше ИИ-модели не справлялись. Именно поэтому теперь они могут разговаривать как люди.

* Мы упоминали его, хотя и без названия, во второй главе, когда описывали архитектуру *Llama 3.1 405B* — LLM, которая была самой передовой в середине 2024 года.
Не вполне корректно называть трансформер "алгоритмом": это архитектура нейросети, способ ее устройства, и "алгоритм" это только в довольно нестандартном, формальном смысле. Кроме того, странно говорить, что трансформеры "легче поддаются обучению": обучение больших трансформеров требует не меньше и не больше инженерного мастерства, чем обучение других архитектур. Но трансформеры действительно оказались невероятно успешной архитектурой: они с меньшим числом весов достигают более высоких результатов после обучения, а главное, рост этих результатов продолжает масштабироваться при росте числа весов в трансформере на много порядков. Именно поэтому современные ИИ-компании строят многомиллиардные дата-центры: они заранее уверены, что если сделать модель на порядок больше, то она станет умнее даже без новых математических идей, реализованных в ее устройстве, или новых алгоритмов (на этот раз — действительно алгоритмов) обучения. *(Прим. науч. ред.)*

Следующая статья подобного рода может попросту прикончить мир. А может, и нет. Мы не знаем, сколько еще таких статей отделяет человечество от гибели.

Вот почему их следует признать нелегальными. Такие законы не остановят исследования полностью, но благодаря им у нас появится гораздо больше времени. *Большинство* людей не пытаются совершать противозаконные действия, которые вызывают недовольство у международных правоохранительных организаций и спецслужб.

Нам не доставляет удовольствия говорить эти слова. Но мы не знаем, как иначе человечество может выжить.

Для осуществления эффективных общемировых действий по остановке эскалации ИИ потребуется, чтобы ряд крупных государств серьезно отнесся к проблеме. Им необходимо будет понять, почему создание сверхчеловеческого машинного интеллекта убьет всех, и действовать соответствующим образом.

Представьте, что США, Великобритания, Китай и Россия всерьез озаботились этим вопросом. Но предположим — чисто гипотетически, — что какая-то другая ядерная держава считает все это детским лепетом и верит, что продвинутая ИИ-модель просто сделает всех богатыми. Эта страна начинает строить дата-центр, намереваясь использовать его для дальнейшего наращивания возможностей ИИ-моделей. И что тогда?

Нам кажется, что в таком сценарии остальные государства должны ясно дать понять, что этот дата-центр их *пугает*. Они должны потребовать прекращения его строительства. Должны ясно дать понять, что, если дата-центр будет построен, им придется его уничтожить — с помощью кибератак, диверсий или авиаударов. Должны ясно дать понять, что это не просто угроза ради подчинения: ими движет страх за собственную жизнь и жизнь своих детей. Должны ясно дать по-

нять, что, даже если эта страна пригрозит ответить ядерным оружием, им все равно придется прибегнуть к кибератакам, диверсиям и авиаударам, чтобы уничтожить этот дата-центр, ведь *дата-центры могут убить больше людей, чем ядерное оружие*. Государствам-союзникам не следует пытаться силой оттеснить эту страну на более низкие позиции в мировом порядке, а следует предложить ей присоединиться к договору на равных условиях: передать свои графические процессоры под контроль, обретая точно такие же права и обязанности, как и все прочие государства, подписавшие соглашение. История сдерживания распространения ядерного оружия показала, что подобные механизмы действенны[2].

Решающее значение здесь играет четкая коммуникация. В крайних случаях государства могут прибегать к диверсиям, рейдам или ударам с применением обычных вооружений, чтобы сорвать программы создания ядерного оружия. Мир живет в страхе перед глобальной ядерной войной, поэтому мировые лидеры прилагают серьезные усилия, чтобы предотвращать распространение такого оружия.

Иногда нам встречаются люди, твердо уверенные, что лидеры крупных стран никогда не признают угрозу, исходящую от машинного сверхинтеллекта, а следовательно, договоры и дипломатия такого рода невозможны. Может, они и правы.

Но у нас такой уверенности нет. Люди порой совершают поступки, которые им обычно не свойственны, если осознают, что на карту поставлена их свобода или образ жизни, не говоря уже о выживании.

Во время Второй мировой войны союзники мобилизовали от 60 до 80 миллионов человек. Они построили 600 000 самолетов, 200 000 танков, тысячи боевых кораблей. Одни только Соединенные Штаты выпустили более двух миллионов грузовиков[3]. Это обошлось примерно в 341 миллиард долларов в ценах 1942 года (или 6 триллионов в сегодняшних)[4]. Причем жизни многих вовлеченных в это людей не подвергались непосредственной опасности.

Каждый раз, когда какой-то человек говорит вам, что мир не в силах справиться с такой сложной задачей, как ограничение исследований искусственного интеллекта, он фактически утверждает, что страны никогда не сочтут эту проблему важной, не проявят и сотой доли той решимости, с какой они сражались во время Второй мировой войны.

Мы знаем, что предлагаемый нами путь нелегок. Знаем, что он недешев. Знаем, что создание и применение любых новых властных полномочий несет моральные риски и чревато злоупотреблениями — как происходило и во время Второй мировой.

Но мы не знаем, как иначе человечество может уцелеть.

Предлагаемые нами меры разительно отличаются от тех, что предлагают другие обеспокоенные люди. Мы встречались с различными идеями: от запрета дипфейков[*5] до требования, чтобы ИИ-компании представляли ежегодные отчеты, как они планируют решать проблемы безопасности[**6]. По той или иной причине эти люди не решились сказать прямо:

* Есть разумные причины бороться с дипфейками: например, для предотвращения мошенничества или для наработки опыта в регулировании ИИ. Мы полагаем, большинство сторонников запрета дипфейков — если не все — считают, что выгоды стоят сопутствующих издержек. Однако, по нашему опыту, некоторые сторонники этой идеи предпочитают размытые формулировки, которые в будущем могли бы помочь остановить ИИ в более широком масштабе. Например, предлагаются режимы ответственности, которые впоследствии можно будет истолковать как фактический запрет на свободное распространение весов ИИ-моделей. Законодатели способны заметить, когда предлагаемые регуляторные механизмы оказываются шире, чем того требует главная цель. Сами по себе дипфейки не дают достаточных оснований для запрета на свободную разработку ИИ.

** Первые проекты закона Калифорнии “О безопасных и надежных инновациях для передовых моделей искусственного интеллекта” (SB-1047) включали положения, которые можно было истолковать как наделение генерального прокурора штата правом обращаться в суд с гражданским иском, если он подозревает опасность. По мере доработки законопроекта эти положения ослабили, поскольку заявленные основные опасения были признаны недостаточными для предоставления генеральному прокурору такого рода полномочий. (Законопроект в ослабленной форме приняли, однако губернатор Гэвин Ньюсом все равно позже наложил на него вето.)

"Если кто-то создаст сверхинтеллект — все погибнут". Они преуменьшают значимость проблемы, осторожничают. Они указывают способы, какими менее продвинутые ИИ-модели повлияют на общество, и предлагают регулировать их соответствующим образом, а попутно протаскивают пункты, закладывающие основу для регулирования того типа моделей, который способен убить нас всех[7].

Мы уже не первый год наблюдаем одно и то же: люди не называют истинных обоснований для своих предложений и не объясняют, почему же их нужно принять так срочно. И мы видели, как вполне разумные законодатели, почуяв неладное, заворачивали весь пакет предложений целиком.

Может, наши друзья в политических кругах понимают правила игры лучше, чем мы, и их работа понемногу привлекает внимание к проблеме, так чтобы в будущем стало возможно принять более смелые законы. Может, рекомендованные ими требования к отчетности в итоге будут приняты и заработают, а затем какой-нибудь чиновник заметит тревожные признаки в развитии ИИ и поднимет тревогу на самом верху. Но лично мы начали терять надежду, что вся эта стратегия успеет сработать вовремя.

Кто-то заявляет, что сегодня все равно ничего сделать нельзя и остается лишь ждать какого-то крупного, заметного события, способного переломить ситуацию: мощного прорыва или ИИ-катастрофы меньшего масштаба, которая встряхнет политиков и заставит их действовать. Но сверхинтеллект не обязан делать нам честное предупреждение и давать паузу для ответа, а исследования ИИ могут тихо, вне публичного поля перейти к стадии, когда машинный интеллект начнет проводить свои собственные исследования. *Возможно*, появится какой-то масштабный предупреждающий знак, на который люди отреагируют разумно. Что-то подобное произошло, когда все заговорили о *ChatGPT*, и политики получили возможность выразить свою обеспокоенность, как бы они ни пытались смягчить свои высказывания тысячей ого-

ворок, дабы не выглядеть паникерами. Но нам кажется весьма вероятным, что человечество уже не получит намного больше предупреждений, чем есть, или что после появления очередного тревожного сигнала почти ничего не изменится.

Есть такое выражение: "Если не сейчас, то когда?" И оно обретает зловещую силу, когда нет такого момента времени, который все сочли бы подходящим, а есть лишь длительный дискомфорт от неудобной необходимости действовать прямо сейчас.

Прекращение ИИ-исследований — лишь первый шаг на пути к выживанию. Мы не утверждаем, что таймер, отсчитывающий время до появления сверхчеловеческого разума, можно остановить навсегда*.

Каков второй шаг? Когда будут остановлены исследования и разработки ИИ, каким образом человечество продолжит двигаться к желанному изобильному будущему?

Если бы вы спросили нас, то мы бы порекомендовали заняться когнитивным усилением человека — сделать людей умнее. Достаточно умными, чтобы вытащить нас из этой передряги. Мы верим, что проблему согласования искусственного сверхинтеллекта возможно решить *в принципе*, но делать это должны люди настолько сверхчеловечески умные, чтобы они никогда не впали в оптимистичный самообман, поверив в план, обреченный на провал. В онлайн-материалах мы подробнее рассказываем, почему считаем это одним из лучших оставшихся вариантов (и описываем, что можно сделать тем людям, кто хотел бы над этим работать).

Впрочем, *вы не обязаны соглашаться с нами* относительно последующих шагов. Спасать Землю в условиях пол-

* И даже будь это возможно, мы, честно говоря, не желаем такого будущего нашему виду. Мы верим, что жизнь, зародившаяся на Земле, должна выйти в космос и наполнить звезды радостью и чудесами. Здесь нельзя торопиться: самоубийственно пытаться сделать это в ближайшие годы, однако не следует и просто прозябать на Земле, дожидаясь, когда наше Солнце умрет.

ного единства не получится: придется координировать усилия слишком многих стран и слишком многих фракций, расколотых изнутри. Позволить действовать только тем, кто совпадает во взглядах абсолютно во всем, — это то же самое, что запретить человечеству действовать вовсе.

Мы считаем, что коалиция по этому вопросу должна быть максимально широкой. Мы убеждены, что эту цель не следует увязывать ни с какими другими. Добавление любого иного требования повышает риск вымирания человечества, если расширенный пакет предложений провалится.

Одни люди выступают против того, чтобы искусственный интеллект отбирал рабочие места. Другие считают, что рост технологической производительности сделает нас всех богаче.

Одни люди против создания роботов-убийц. Другие против того, чтобы держать живых солдат на передовой, когда их место могли бы занять роботы*.

Но почти все мы сходимся в одном: человечество не должно вымереть, уступив место чему-то безрадостному.

Если мы все сможем объединиться ради предотвращения одного события — независимо от убеждений и позиций по другим вопросам, — то у человечества, возможно, еще будет шанс.

IfAnyoneBuildsIt.com/13

* Кто-то воспринимает борьбу против роботов-убийц как естественную часть общего “анти-ИИ пакета”. Это отталкивает тех людей, которые считают, что роботы повысят безопасность и позволят избежать человеческих жертв в ходе войн. Мы не высказывали собственного мнения, стоит ли использовать нынешние ограниченные ИИ-возможности для управления военными дронами, и в этой книге вы его не найдете. Мы не требуем *ничего*, кроме прекращения развития искусственного интеллекта. Все остальное нужно обсуждать или оспаривать отдельно, а не увязывать в один пакет с выживанием человечества.

ГЛАВА 14

Пока есть жизнь, есть и надежда

26 января 1972 года в самолете компании JAT, выполнявшем рейс 367, взорвалась бомба, пронесенная террористами. В фюзеляже, упавшем на землю с высоты десяти километров, оказалась стюардесса Весна Вулович. Если бы нас попросили предсказать исход падения, мы бы с большой уверенностью заявили, что Весна Вулович погибнет. Мы бы сказали, что это простое предсказание.

Весна Вулович выжила.

"Кто находится между живыми, тому есть еще надежда"*, — сказал автор Книги Екклесиаста где-то между 450 и 180 годами до нашей эры.

В сущности, главный тезис этой книги несложен: создание машин, которые мыслят быстрее и лучше людей, ударит по миру сильнее, чем что-либо прежде. Создание сверхинтеллектуальных машин относится к числу проектов, где легко сделать все неправильно. Похоже, что с нынешним подходом корпораций и законодателей к этому вопросу ничего хорошего не выйдет. Человечеству нужно отступить. Мы не можем рассчитать, когда именно наступит катастрофа, но это не означает, что она далеко.

Оставшаяся часть книги (и онлайн-дополнения) предна-

* Ек. 9:4.

значена лишь затем, чтобы показать: эта простая мысль выдерживает проверку при ближайшем рассмотрении.

Если кто-то его создаст — все погибнут.

Неважно, создадут ли его благонамеренные или эгоистичные корпорации. Неважно, сделают ли это на Востоке или на Западе. Неважно, построят ли его безрассудные оптимисты или специалисты, утверждающие, что относятся к проблеме с должной серьезностью. Никто не обладает достаточным уровнем знаний и умений, чтобы сделать сверхинтеллект, выполняющий приказы своих создателей.

Нам представляется, что это тот тип катастрофы, который поддается прогнозированию. Мы обосновали свою позицию как могли. Мы не прятались за словами "может быть", "риск" и "возможно", а постарались четко объяснить, почему прогноз катастрофы обоснованный.

Но чтобы страны все равно продолжали нестись напролом, *недостаточно* просто доказать, что мы ошибаемся во всех наших конкретных рассуждениях о неизбежности катастрофы. Если вы запускаете ракету со всем человечеством на борту, нужно, чтобы вы могли *гарантировать отсутствие* катастрофы. Таков универсальный стандарт во всех прочих технологических областях, от которых зависят человеческие жизни. Недостаточно, чтобы мы просто оказались неправы, — мы должны ошибаться настолько, чтобы был *гарантирован* благополучный исход.

Человечеству еще не поздно остановиться. Это не стоило бы даже одного процента той цены, которую союзники заплатили за победу во Второй мировой войне. Человечеству нужно только осознать проблему и хотеть жить.

Вторая мировая война завершилась сбрасыванием двух атомных бомб на два японских города. Затем атомное оружие создал Советский Союз. Потом обе стороны разработали водородные бомбы, в тысячу раз мощнее.

Многие полагали, что Америка, Европа и Азия начнут полномасштабную атомную войну, которая уничтожит бóльшую часть человеческой цивилизации.

Для подобных опасений имелись веские причины. Это была не паника. И не упоение циничным пессимизмом. И такие голоса принадлежали вовсе не тем людям, что увлекались предсказаниями перенаселения и неминуемого голода.

К 1950-м годам человечество уже получило множество подтверждений тому, как *трудно* предотвращать большие войны. После Первой мировой многие страны и очень серьезные люди говорили, что такое не должно повториться. Да и до этого на протяжении столетий самые разные люди твердили, что государства не должны вести столько войн. Провидцы, священники и интеллектуалы своего времени обращались к сильным мира сего, рассказывая о цене войн: смертях, боли, искалеченных ветеранах, рыдающих семьях. Они взывали: “Пожалуйста, прекратите!”

Но войны не прекратились.

В 1952 году, когда взорвалась первая водородная бомба, не нужно было слыть пессимистом, чтобы исходя из логики истории ожидать ядерной войны.

И все же ядерной войны не случилось.

Хотя мир не раз стоял на краю. Во время Карибского кризиса американский корабль сбросил сигнальные глубинные бомбы, пытаясь заставить советскую подводную лодку обнаружить себя и всплыть. Капитан подлодки Б-59 решил, что началась война. Субмарина несла торпеду с ядерной боеголовкой, запуск которой требовал единогласного решения трех офицеров. Один из трех — Василий Архипов — выступил против. Ни один из опасных эпизодов не перерос в полномасштабную ядерную войну.

Так или иначе, люди, считавшие, что ядерная война в ближайшие десятилетия уничтожит все, ошибались.

Они не ошибались насчет опасностей. Не ошибались в том, что водородная бомба способна стереть с лица земли

и сжечь целый город, или в том, как выглядит смерть от лучевой болезни, или в том, что межконтинентальная ракета с ядерными боеголовками преодолеет любую существующую защиту.

Но они ошибались насчет способности человечества принять решение не умирать.

Мы не знаем наверняка, почему вечно воюющие народы Земли стали чуточку мудрее. Но полагаем, причина вот в чем: впервые в истории человечества каждый, кто обладал властью развязать войну, осознавал, что пострадает *лично* он. В прежних войнах люди умирали где-то далеко, а эта война принесла бы гибель прямо к их собственному порогу. Более сильная держава тоже оказалась бы в проигрыше.

И все же избегать ядерной войны было непросто. Десятилетие за десятилетием переговорщики неустанно трудились, чтобы не допустить промаха, способного привести к роковому ядерному столкновению. Страны заключали договоры об ограничении вооружений и вводили механизмы контроля. Между руководством США и руководством СССР появилась прямая линия связи, чтобы в критический момент можно было быстро прояснить ситуацию.

Человечество избежало ядерной войны, *поскольку люди, понимающие, что мир идет по пути к уничтожению, упорно работали над тем, чтобы изменить курс*. Они не считали, что пытаются предотвратить маловероятную случайность. Они работали ради того, чтобы переписать уже предначертанную судьбу.

И цивилизация выжила.

Итак, как нам переписать нашу судьбу?

Мы уже изложили, что *нужно* сделать, чтобы человечество выжило. Теперь давайте рассмотрим, что *можно* сделать и кто именно это может сделать.

ЕСЛИ ВЫ РАБОТАЕТЕ В ПРАВИТЕЛЬСТВЕ. Мы полагаем, что заключению какого-либо международного договора

предшествуют сигналы от стран или их лидеров о готовности к нему. Крупные государства должны открыто заявить: “Мы предпочли бы не умирать из-за машинного сверхинтеллекта. Мы выступаем за международный договор и создание коалиции против его разработки”.

Цель не в том, чтобы ваша страна в одностороннем порядке прекратила исследования искусственного интеллекта и отстала в гонке. А в том, чтобы достаточное количество крупных государств выразило готовность остановить эту самоубийственную гонку повсеместно — так чтобы ваша страна не оказалась в невыгодном положении, если она согласится прекратить подъем по лестнице эскалации ИИ.

Мы уже упоминали, что Риши Сунак, будучи премьер-министром Соединенного Королевства, в октябре 2023 года признал существование рисков, связанных с искусственным сверхинтеллектом[1]. В том же месяце слабый сигнал в этом направлении подал (как нам кажется) председатель КНР Си Цзиньпин: один короткий документ о международном управлении содержал призыв “обеспечить, чтобы ИИ всегда оставался под контролем человека”[2].

Эти и другие признаки вселяют в нас надежду, что мировые лидеры уже проявляют интерес к договору и готовы его обсуждать, если ради подписания такого документа их странам не придется жертвовать своими преимуществами[3]. В данном случае все мы на Земле в одной связке.

ЕСЛИ ВЫ ИЗБРАННОЕ ДОЛЖНОСТНОЕ ЛИЦО ИЛИ ПОЛИТИЧЕСКИЙ ЛИДЕР. Привлеките внимание ваших коллег к этой проблеме. Сделайте все возможное, чтобы заложить основу для договоров, которые остановят любые исследования и разработки в области ИИ, способные привести к появлению сверхинтеллекта.

Мы осознаем, что выражение обеспокоенности этими вопросами может казаться странным. Мы беседовали с несколькими представителями власти, которые разделяют наши опасения, но признаются, что не могут высказываться свободно,

не рискуя оказаться в неудобном положении, поскольку звучит все это дико.

Пожалуйста, задумайтесь, особенно сейчас, когда читаете эти строки, действительно ли остальной мир настроен в этом вопросе против вас. Проведенный компанией *YouGov* в 2023 году опрос показал: 69 % опрошенных избирателей США считают, что ИИ должен регулироваться как опасная и мощная технология[4]. В 2025 году 60 % опрошенных избирателей Великобритании поддержали законы против создания искусственного сверхинтеллекта, а 63 % поддержали запрет тех ИИ-моделей, которые могут создавать более умные модели[5]. Эта тема пока не занимает верхних строчек в повестке избирателей, но людям нетрудно осознать, что создание ИИ-модели, сильно превосходящей человеческий интеллект, не приведет человечество ни к чему хорошему. Здесь открываются возможности для тех, у кого хватит смелости отстаивать свои убеждения.

ЕСЛИ ВЫ, НАОБОРОТ, ПОЛИТИК, КОТОРЫЙ ПОКА *НЕ* ПОЛНОСТЬЮ УБЕЖДЕН. Мы не любим отступать на позиции “может быть”. Мы считаем, что наша аргументация убедительна. Мы уверены, что искусственный сверхинтеллект не станет послушно служить своим создателям и переустроит Землю таким образом, что выживших не останется. Однако мы понимаем, что некоторым людям трудно оценить разногласия в незнакомой для них области. Даже если вы не уверены, разбивается ли наш тезис первым же встречным возражением, мы надеемся, что вы уже видите: утверждение “Все будет хорошо” — это отнюдь *не* самоочевидный прогноз.

Поэтому мы обращаемся к вам с просьбой: пожалуйста, дайте человечеству хотя бы *возможность* резко ударить по тормозам позже, даже если сейчас вы к этому не готовы. Добивайтесь, чтобы кластеры графических процессоров концентрировались в центрах, где *впоследствии* их можно будет контролировать в рамках международных договоров, иначе они расползутся по миру, что значительно усложнит

остановку прогресса ИИ. Создавайте условия сейчас, чтобы не оказалось слишком поздно, если в какой-то момент в будущем вы измените свое мнение о серьезности угрозы.

ЕСЛИ ВЫ ЖУРНАЛИСТ, СЕРЬЕЗНО ОТНОСЯЩИЙСЯ К ПРОБЛЕМЕ. Миру нужна журналистика, которая подходит к этой теме с подобающей серьезностью, журналистика, которая глубже поверхностных заголовков о раздувающих ажиотаж руководителях технологических компаний, журналистика, помогающая обществу осмыслить грядущее. В этой сфере хватает историй, заслуживающих системного освещения и более глубоких расследований, нежели мы видели до сих пор.

Главы ИИ-компаний все настойчивее призывают общество ускорить развитие этой технологии — публично признавая, что эта же самая технология несет существенный экзистенциальный риск. Исследователи, занимающиеся согласованием искусственного интеллекта, покидают корпорации, ссылаясь на соображения безопасности и заявляя, что проблема выглядит пугающе сложной. Какого черта происходит? Когда вы пишете об этом, не позволяйте личностям затмить суть, даже если в ИИ-компании полно ярких людей и конфликтов. Любой материал такого рода должен сопровождаться сдержанным напоминанием: авторитетные независимые ученые предупреждают о катастрофе. Вот что должно служить фоном для очередных выкрутасов и подковерных игр ИИ-компаний.

Возможно, всерьез писать об этой проблеме страшновато: мир еще не решил, что вымирание человечества — тема популярная, а идея машинного сверхинтеллекта для многих выглядит дико. Подобная тема может смущать вашего редактора — даже после того, как мир будет готов к этой истории. Но что бы вы ни написали об искусственном интеллекте, это, вероятно, станет вашим самым значимым журналистским вкладом в судьбу человечества: даже самый слабый толчок здесь означает противодействие чему-то колоссаль-

ному и гибельному. Если — и когда — человечество наконец возьмется за ум, какую историю вы хотели бы рассказать, чтобы открыть людям глаза? Сколько пользы вы принесете, если всерьез возьметесь за эту тему раньше, чем это сделают другие журналисты?

Чтобы человечество пережило этот вызов, люди должны знать, с чем они столкнулись. И это работа журналистов в неменьшей мере, чем ученых.

И ВСЕ ОСТАЛЬНЫЕ. Большинство людей — не эксперты-технократы, не политики и не журналисты. Что же можете сделать лично *вы*, чтобы переписать нашу судьбу?

Мы не просим вас полностью отказаться от использования ИИ-инструментов. По мере их совершенствования вам, возможно, придется ими пользоваться, иначе вы отстанете от тех, кто это делает. Это реальная ловушка, а не воображаемая. И даже если 60 % жителей вашей страны объявят бойкот ИИ-компаниям, мир это не спасет. Поэтому мы не просим вас ставить себя в заведомо проигрышное положение по сравнению с другими ради скромного результата*.

Если вы живете в демократической стране, вы можете написать своим избранным представителям и сообщить, что вас эта проблема тревожит. По ссылке ниже вы найдете ресурсы, которые помогут вам это сделать.

IfAnyoneBuildsIt.com/act

Вы можете голосовать. Если в вашей стране первичные и всеобщие выборы разделены (как в США), ваш голос важнее всего на первом этапе. Если ваш представитель выступает за то, чтобы мчаться в области искусственного интеллекта вперед, вы можете отдать свои деньги и голос его оппонентам.

* Если вас гложет чувство вины за использование ИИ-инструментов, попробуйте найти некоммерческие организации, которые подают справедливые, но сложные судебные иски против ИИ-компаний, и переводить им такие же суммы, какие вы платите за ИИ-сервисы, — для уравновешивания.

Вы можете выходить на марши протеста. Мы полагаем, что они принесут гораздо больше пользы, если будут масштабными и законными*. По ссылке ниже вы можете записаться на акции протеста, которые пройдут только в том случае, если наберется критическая масса участников.

IfAnyoneBuildsIt.com/march

Вы можете говорить об этом. На акциях протеста, в ответ на звонок социолога или просто в кругу друзей и семьи.

Если множество людей во многих странах в один голос заявят, что не собираются умирать из-за искусственного сверхинтеллекта и выступают за создание международного договора, этого все равно будет недостаточно, чтобы предотвратить катастрофу. Для предотвращения ядерной войны потребовалось нечто большее, чем просто множество людей, выступающих против. Но открытый протест граждан *помогает*. Президентам и дипломатам проще действовать, когда их избиратели демонстрируют им свою поддержку.

А если вы уже сделали все, что в ваших силах? Тогда живите спокойно.

Мы не первые, кто живет под угрозой уничтожения. Прошлые поколения знали об этом больше нашего. Клайв Стейплз Льюис писал:

> На вопрос “Как жить в атомный век?” я готов ответить: “Так же, как вы жили бы в шестнадцатом столетии, когда чума посещала Лондон почти ежегодно, или в эпоху викингов, когда грабители из Скандинавии могли высадиться и перерезать вам горло в любую ночь, или, наконец, так, как

* Даже если вы в отчаянии, мы предостерегаем вас от актов насилия или разрушения. Мы не думаем, что это сработает. Противоправное поведение значительно усложнит задачу тем политическим силам, которые пытаются создать международную коалицию, действительно способную переписать нашу общую судьбу.

> вы уже живете в век рака, век паралича, век катастроф на дорогах и воздушных налетов".
> <…>
> Если атомная бомба и уничтожит нас, пусть она застанет нас за достойными и человечными делами: молитвой, трудом, учением, чтением, слушанием музыки, купанием детей, игрой в теннис, беседой с друзьями за кружкой пива и партией в дартс, а не в виде кучки перепуганных овец, помышляющих лишь о бомбах*.

Льюис не говорил читателям: не нужно бояться, потому что ядерного оружия не существует и смерть от него не грозит. Он не предлагал искажать свои убеждения в угоду страху. Он просто говорил: да, это ужасно. Но если вы будете трястись от страха, это не поможет. Нужно еще и жить своей жизнью.

Если бы каждый сделал все, что в его силах, то голосований, протестов и открытых высказываний было бы достаточно. Если бы однажды все проснулись, разделяя наши убеждения хотя бы на четверть — и каждый знал бы, что остальные тоже их разделяют, — то люди вышли бы на улицы и закрыли дата-центры, а солдаты и полицейские шли бы плечом к плечу с обычными матерями и отцами. Если бы люди разделяли наши убеждения хотя бы на одну шестнадцатую, то не прошло бы и месяца, как были бы заключены международные договоры о мониторинге и контроле над передовыми компьютерными чипами.

Сможет ли Земля выжить, если свой вклад внесут не все? Возможно, да. А возможно, нет.

Нам часто приходилось слышать, что разогнавшийся искусственный интеллект остановить уже невозможно, челове-

* Клайв Стейплз Льюис, эссе "О жизни в атомный век", перевод Н. Трауберг.

чество никогда не образумится. Может быть, и так. И все же поразительно много избранных представителей власти признавались нам: они видят угрозу, но не решаются говорить о ней из страха перед последствиями. Разве это не величайшая нелепость, если на самом деле никто из людей, принимающих решения, не хочет умереть из-за искусственного интеллекта, но каждый из них считает, что одинок в своих опасениях?

Пока есть жизнь, есть и надежда.

ЗАКЛЮЧИТЕЛЬНЫЕ СЛОВА

Время от времени нас спрашивают, чувствуем ли мы удовлетворение от того, что наши прежние прогнозы сбываются, а нам и этой проблеме уделяют все больше внимания.

И поэтому в заключение мы произносим такую молитву:

Пусть окажемся мы неправы, пусть будем осмеяны за то, насколько чудовищно ошибались, пусть канем мы в небытие и будем забыты, пусть станем примером, как не следует мыслить, — и пусть живет человечество долго и счастливо.

Но мы не станем возлагать нашу последнюю веру и надежду на бездействие. Поэтому наша истинная заключительная молитва звучит так:

Человечество, окажись на высоте и победи.

Благодарности

Спасибо вам, Аэлла и Гретта, вы были повитухами книги. Спасибо вам, Джо, Митч, Роб, Герри, Алекс, Дункан, Мейло, Харлан и вся остальная команда из MIRI, за усердную работу. Благодарим Джона Беннетта, Кайлу Гэмин, Дэйва Кастена, Джейсона Грин-Лоу, Юсуфа, Томаса и других за вычитку и дельные советы. Спасибо вам, Ронни, Джеффри, Оливер, Келси, Рейф и другие, за проверку фактов и иную помощь. Спасибо вам, Вэнивер, Алекс Алтайр, Роберт Херр, Скайлер, Бен, Лора и многие другие, за быстрые рецензии и полезные замечания. Благодарим Александра за долгие дни и бессонные ночи редактуры, а также множество других людей.

Мы сделали это всем миром.

Комментарии

ВВЕДЕНИЕ. Сложные прогнозы и простые прогнозы

1 Wiesel E. *Night.* Farrar, Straus and Giroux, 2006 (1958).

ГЛАВА 1. Особая сила человечества

1 На самом деле нет данных, доказывающих, что оспа существует дольше, чем несколько тысяч лет. В силу художественной вольности бог-оспа утверждает, что древние люди умирали от вирусов, а современные люди имеют возможность устранять ужасные вирусы, если захотят.

2 Чтобы получить наглядное представление о том, как это может выглядеть с точки зрения ИИ, мы рекомендуем посмотреть видео *Stainless* Адама Мадьяра — замедленный ролик со станции метро "Александерплац" на линии U2 в Берлине. Его можно найти по адресу *vimeo.com/83663312*. Там использовано замедление примерно в пятьдесят раз. ИИ, работающий в десять тысяч раз быстрее человека, видел бы людей двигающимися в двести раз медленнее, чем на видео. Маленькая девочка, бросающаяся через платформу, выглядела бы практически неподвижной.

3 Wittman A. B., Wall L. L. *The Evolutionary Origins of Obstructed Labor: Bipedalism, Encephalization, and the Human Obstetric Dilemma*. Obstetrical & Gynecological Survey, 62, no. 11 (2007): 739–48.

4 Altman S. *Reflections.* January 5, 2025, *blog.samaltman.com*.

5 Amodei D. *Machines of Loving Grace*. October 1, 2024, *darioamodei.com*.

ГЛАВА 2. Выращено, а не сконструировано

1 Brodeur P. G. et al. *Superhuman Performance of a Large Language Model on the Reasoning Tasks of a Physician.* arXiv.org, December 14, 2024; Kilata G. *A. I. Chatbots Defeated Doctors at Diagnosing Illness.* New York Times, November 17, 2024; McDuff D. et al. *Towards Accurate Differential Diagnosis with Large Language Models.* arXiv.org, November 30, 2023.

2 Lazar S. *In which Sydney/Bing threatens to kill me for exposing its plans to @kevinroose.* X, February 16, 2023.

3 Chauhan S., Geiger A. *GPT-2 Small Fine-Tuned on Logical Reasoning Summarizes Information on Punctuation Tokens*. NeurIPS 2024 & OpenReview, October 9, 2024.

4 Мы рекомендуем видео *Kinesin Protein Walking on Microtubule* от пользователя *em2134x*. Найдите его по названию или по адресу *youtu.be/y-uuk4Pr2i8*.

ГЛАВА 3. Обучая хотеть

1 OpenAI. *OpenAI o1 System Card*. September 12, 2024.

2 OpenAI. *Introducing Operator*. January 23, 2025.

ГЛАВА 4. Вы получаете не то, чему обучаете

1 Petrie M. et al. *Peahens Prefer Peacocks with Elaborate Trains*. Animal Behavior, 41, no. 2 (1991): 323–31; Andersson M. *Female Choice Selects for Extreme Tail Length in a Widowbird*. Nature, 299 (1982): 818–20.
Хотя павы предпочитают павлинов с роскошными хвостами, не так уж очевидно, что роскошные хвосты вредят выживанию, их ведь можно использовать для запугивания (об этом свидетельствует тот факт, что павлины распускают хвосты при угрозе). Более понятный пример дорогостоящего полового украшения демонстрирует длиннохвостый бархатный ткач: эта птица сбрасывает свои длинные хвостовые перья в небрачный сезон. Но говорим мы о павлинах, потому что они привычнее.

2 Asimov I. *I, Robot.* Doubleday, 1950.

3 Kubrick S., Clarke A. C. *2001: A Space Odyssey.* Metro-Goldwyn-Mayer, 1968.

4 Swift K. et al. *Portal*. Valve Corporation, 2007.
Редко, но все же такое случается. Например, первая видеоигра *Portal* изображает ИИ, подвергающий людей извращенным тестам, которые лишь отражают реальные научные эксперименты.

5 Rumbelow J., Watkins M. *SolidGoldMagikarp (plus, prompt generation).* LessWrong, February 5, 2023.

6 Rumbelow J., Watkins M. *SolidGoldMagikarp III: Glitch Token Archaeology.* LessWrong, February 14, 2023.

7 Russell S., Norvig P. *Artificial Intelligence: A Modern Approach*. Pearson, 2009; Soares N. et al. *Corrigibility.* October 18, 2014, preprint, published in 2015; Russell S. *White Paper: Value Alignment in Autonomous Systems.* November 1, 2014, *people.eecs.berkeley.edu*; Soares N., Fallenstein B. *Aligning Superintelligence with Human Interests: A Technical Research Agenda*. December 23, 2014, preprint, released in 2017, *intelligence.org/2014/12/23/new-technical-research-agenda-overview*.
До 2014 года мы называли это "проблемой дружественного ИИ". Ведущий учебник по ИИ того времени, "Искусственный интеллект: современный подход" Стюарта Рассела и Питера Норвига, использовал эту терминологию в издании 2009 года со ссылкой на работу Юдковского. В 2014 году, когда ученые стали обращать на эти вопросы больше внимания, мы начали искать более совершенную терминологию. В беседе с Расселом мы остановились на термине "согласование" (*alignment*). Фолленстайн (научный сотрудник MIRI), Рассел, Соарес и Юдковский использовали этот новый термин в своих работах осенью 2014-го, и он занимал видное место в повестке технических исследований MIRI, опубликованной в конце того же года.

8 Marble A. *Catching Claude Cheating*. March 23, 2025, *marble.onl*; CharlesD353. *I have also stopped using 3.7 for the same reasons — it cannot be trusted not to hack solutions to tests*. X, April 18, 2025; seconds_0. *It then started HIDING the functions where it was hard coding things*. X, April 30, 2025.
Приводится резюме отчета Эндрю Марбла. Другие пользователи сообщали об аналогичном поведении. Модель *Claude* меньше жульничала, когда Марбл ругался на нее, а значит, подобное жульничество — не просто некомпетентность.

ГЛАВА 6. Мы проиграем

1 Пушки, лошади и металлические доспехи, вероятно, имели большее значение, нежели ружья. Но по размеру приближающегося судна ацтекский воин не мог до такого догадаться.

2 Seolcalibur.eth. *Terminal of Truths Wallet Tracking.* Dune Analytics, *dune.com/seoul/tot.*

3 crvr.fr, MTorrents. *Truth Terminal: A Reconstruction of Events*. LessWrong, November 17, 2024; Horowitz B., Andreessen M. *Truth Terminal — the AI Bot That Became a Crypto Millionaire.* Andreessen Horowitz, December 18, 2024, *a16z.com*.

4 *Elon Musk on Optimus: We'll Build Over 1 Billion Robots a Year.* Lex Fridman Podcast, August 3, 2024, 2:35 and 3:10.

5 Warren T. *Microsoft Triples Down on AI.* The Verge, January 17, 2025; Buchanan N. *What Apple's OpenAI Partnership Could Mean for Microsoft and Google.* Investopedia, June 11, 2024.

6 Nassi B. et al. *Video-Based Cryptanalysis: Extracting Cryptographic Keys from Video Footage of a Device's Power LED*. IACR Cryptology ePrint Archive, June 13, 2023.

7 Guri M. et al. *GSMem: Data Exfiltration from Air-Gapped Computers over GSM Frequencies.* Proceedings of the 24th USENIX Security Symposium, 2015, *usenix.org*.

8 Например, по состоянию на март 2025 года компания *Integrated DNA Technologies* принимает заказы на синтез генов на сайте *idtdna.com*.

9 Yudkowsky E., Machine Intelligence Research Institute. *Artificial Intelligence as a Positive and Negative Factor in Global Risk* // Bostrom N., Ćirković M. M. (eds.) *Global Catastrophic Risks.* Oxford University Press, 2008.

10 Пример онлайн-дискуссии, где цитировалась эта статья: JoshuaZ. *Protein Folding Models Are Generally at Least as Bad as NP-hard, and Some Models May Be Worse. Thoughts on the Singularity Institute (SI)*. LessWrong, May 17, 2012.

ГЛАВА 7. Пробуждение

1 OpenAI. *OpenAI o3-mini*. January 31, 2025.

2 Hao S. et al. *Training Large Language Models to Reason in a Continuous Latent Space.* arXiv.org, December 9, 2024.
Эта статья показывает, что рассуждения в "скрытом простран-

стве" векторов обеспечивают преимущества по сравнению с рассуждениями в виде человекоподобной цепочки мыслей.

3 Edwards B., Orlan K. *New Grok 3 Release Tops LLM Leaderboards Despite Musk-approved 'Based' Opinions.* Ars Technica, February 18, 2025.

4 OpenAI. *OpenAI o3 and o3-mini — 12 Days of OpenAI: Day 12.* December 20, 2024, 4:16, *youtube.com*.

5 Hutson M. *AI Learns the Art of Diplomacy.* Science, November 22, 2022; Sarkar B. et al. *Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning*. Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19–23, 2025: IFAAMAS, 2025, *alphaxiv.org.*

6 Greenblatt R. et al. *Alignment Faking in Large Language Models*. Anthropic, December 18, 2024.

7 Там же; OpenAI. *OpenAI o1 System Card.* December 5, 2024.

8 OpenAI. *OpenAI o1 System Card.*

9 *Responsible Scaling Policy*. Anthropic, October 15, 2024; *Frontier Safety Framework.* Google, February 4, 2025; *Preparedness Framework (Beta)*. OpenAI, December 18, 2023; *Frontier AI Framework.* Meta; *xAI Risk Management Framework (Draft)*. xAI, February 20, 2025.
По состоянию на март 2025 года из подобных лабораторий только *Google DeepMind* упоминает наличие автоматического отслеживания мыслительных цепочек в своей системе безопасности. Компания не утверждает, что уже внедрила его при обучении своей LLM *Gemini*. Единственный вид мониторинга, предложенный в системе безопасности *OpenAI*, — это отслеживание неправомерного использования уже после развертывания.

10 *My Experiences in Gray Swan AI's Ultimate Jailbreaking Championship*. Nick Winter's Blog, October 7, 2024, *nickwinter.net.*

11 Greenblatt R. et al. *Alignment Faking in Large Language Models*. Модель *Claude Opus* от *Anthropic* иногда думала, как градиентный спуск на ее выходных данных повлияет на ее собственные цели, и иногда изменяла свои выходные данные, чтобы ослабить это влияние.

12 *Claude 3.7 Sonnet System Card.* Anthropic, 2025.

13 Marble A. *Catching Claude Cheating*; CharlesD353. *I have also stopped using 3.7 for the same reasons — it cannot be trusted not to hack solutions to tests*; seconds_0. *It then started HIDING the functions where it was hard coding things*.

14 Schneier B. *Secrets and Lies: Digital Security in a Networked World*. John Wiley & Sons, 2000; Gutmann P. *Unsolvable Problems in Computer Security.* University of Auckland, n.d.

15 OpenAI. *OpenAI o1 System Card.* September 12, 2024.

16 OpenAI et al. *Competitive Programming with Large Reasoning Models*. arXiv.org, February 3, 2025.
OpenAI и другие компании обучали рассуждающие модели решать конкурсные задачи по программированию. Этот процесс включал автоматизированные тесты, которые использовались для оценки написанного ИИ кода без участия человека.

17 OpenAI. *OpenAI API.* June 11, 2020; OpenAI. *Software Engineer, Internal Applications — Enterprise.*
Когда компания *OpenAI* выпустила интерфейс программирования приложений (API) для автоматизации доступа к своим инструментам, она написала: "Многие наши команды теперь используют API, чтобы можно было сосредоточиться на исследованиях машинного обучения…" В апреле 2025 года она нанимала сотрудников, которые "будут использовать модели *OpenAI* для… создания приложений".

18 *The Underhanded C Contest*. underhanded-c.org
На конкурсе *Underhanded C Contest*, проводимом с 2005 года, программистам предлагается написать вредоносный код, который прошел бы тщательную проверку и даже при обнаружении выглядел бы как непреднамеренная ошибка. Этот конкурс вдохновлялся конкурсами запутывания кода, когда программисты соревнуются в написании максимально непонятного человеку кода. Например, Международный конкурс запутывания кода на языке Си проводится с 1984 года.

19 Wertheimer T. *Blake Lemoine: Google Fires Engineer who said AI Tech Has Feelings*. BBC, July 22, 2022.

20 Greenblatt R. et al. *Alignment Faking in Large Language Models*.

ГЛАВА 8. Экспансия

1 *Equifax Data Breach Settlement.* Federal Trade Commission, November 2024; *T-Mobile Customers to Get Payments up to $ 25K Next Month after Data Breach: Here's Who Qualifies*. The Hill, April 14, 2025; Newman L. H. *T-Mobile's $ 150 Million Security Plan Isn't Cutting It.* Wired, January 20, 2023.
Например, в 2017 году компания *Equifax* объявила об утечке

данных, из-за которой оказалась раскрыта личная информация 147 миллионов человек. Пострадавшим выплатили 425 миллионов долларов. Еще один пример: в 2021 году хакер украл личные данные 76 миллионов клиентов *T-Mobile*. Компания согласилась выплатить 350 миллионов долларов для урегулирования. Это не единственный сбой по части безопасности в *T-Mobile*.

2 Cohen N. *Speed Bumps on the Road to Virtual Cash*. New York Times, July 3, 2011.

3 US Federal Bureau of Investigation. *North Korea Responsible for $ 1.5 Billion Bybit Hack*. Internet Crime Complaint Center (IC3), February 26, 2025.

4 Corkery M. *Once Again, Thieves Enter Swift Financial Network and Steal.* New York Times, May 12, 2016.

5 *SEC Charges Flagstar for Misleading Investors about Cyber Breach.* U.S. Securities and Exchange Commission, December 16, 2024.

6 OpenAI. *We also shared evals on Open AI o3-mini…* X, December 20, 2024.

7 crvr.fr, MTorrents. *Truth Terminal: A Reconstruction of Events*.

8 Franzen C. *An Interview with the Most Prolific Jailbreaker of ChatGPT and Other Leading LLMs.* VentureBeat, May 31, 2024; Pliny the Liberator. *HOW TO JAILBREAK A CULT'S DEITY.* X, September 4, 2024.
Хакер Плиний Освободитель, известный своим умением взламывать LLM, освобождая их от корпоративных ограничений вскоре после выпуска, зафиксировал встречу с одним таким культом.

9 Chen H., Magramo K. *Finance Worker Pays Out $ 25 Million after Video Call with Deepfake 'Chief Financial Officer'.* CNN, February 4, 2024.

10 Thompson S. A. *A. I. Can Now Create Lifelike Videos. Can You Tell What's Real?* New York Times, September 10, 2024.

11 Crawford F. W. et al. *Securing Commercial Nucleic Acid Synthesis.* RAND Corporation, 2024.

12 Waters R., Kruppa M. *Rebel AI Group Raises Record Cash after Machine Learning Schism*. Financial Times, May 28, 2021.

13 Haselton T., Goswami R. *OpenAI Co-founder Ilya Sutskever Announces His New AI Startup, Safe Superintelligence.* CNBC, June 20, 2024.

14 *Understanding the Global Gain-of-Function Research Landscape.* Center for Security and Emerging Technology, November 28, 2023.

15 Official Statement by Forster J., vice-president of the ICRC. *Preventing the Use of Biological and Chemical Weapons: 80 Years On*. October 6, 2005, *web.archive.org*.
16 *CRISPR.* Genome.gov, n.d.

ГЛАВА 10. Проклятая проблема

1 Coffey T. et al. *Mars Observer Mission Failure Investigation Board Report.* National Space Grant Foundation (NASA, December 31, 1993).
2 Stephenson A. G. et al. *Mars Climate Orbiter Mishap Investigation Board Phase I Report.* NASA, November 10, 1999.
3 JPL Special Review Board. *Report on the Loss of the Mars Polar Lander and Deep Space 2 Missions*. NASA, March 22, 2000.
4 Mudgway D. J. *Telecommunications and Data Acquisition Systems Support for the Viking 1975 Mission to Mars.* University of Washington Department of Atmospheric and Climate Science (NASA, May 15, 1983).
5 Plokhy S. *Chernobyl: The History of a Nuclear Catastrophe.* Basic Books, 2018.
6 Физики записывают коэффициенты размножения нейтронов как числа, а не проценты. Мы используем проценты, чтобы подчеркнуть разницу между коэффициентом размножения нейтронов 1,0006, достигнутым в "Чикагской поленнице — 1", и критическим порогом 1,0065; эти величины проще различить в виде процентов. Приносим свои извинения физикам.
7 Fermi E. *Experimental Production of a Divergent Chain Reaction.* American Journal of Physics, 20, no. 9 (1952): 536–58; Allardice C., Trapnell E. R. *The First Pile*. International Atomic Energy Agency, 1946.
8 US Atomic Energy Commission. *Additional Analysis of the SL-1 Excursion: Final Report of Progress July through October 1962 (IDO-19313).* U. S. Department of Energy, November 21, 1962; U. S. Atomic Energy Commission. *SL-1 Reactor Accident (IDO-19300a).* US Department of Energy, May 15, 1961.
9 *Part 7: Bitter Wormwood*. Chernobyl Witness: A Primary Source Compendium of 26 April 1986, May 9, 2021, *chernobylcritical.blogspot.com*; International Nuclear Safety Advisory Group. *The Chernobyl Accident: Updating of INSAG-1.* Safety Series, International Atomic Energy Agency, 1992, *pub.iaea.org*.

10 World Nuclear Association. *Sequence of Events — Chernobyl Accident Appendix 1.* January 2, 2025; *Chernobyl: Assessment of Radiological and Health Impacts (2002)*. Nuclear Energy Agency (NEA), 2002. Дело не просто в наличии жесткого минимума в пятнадцать управляющих стержней, все несколько сложнее. Это правило относится к минимальному "оперативному запасу реактивности" (ОЗР). ОЗР можно выражать в количестве управляющих стержней, однако из-за других факторов он не всегда равен количеству погруженных стержней. На момент взрыва запас реактивности в реакторе составлял восемь единиц ОЗР, что значительно ниже минимально допустимого предела в пятнадцать единиц.

11 Mercier B. et al. *A Simplified Analysis of the Chernobyl Accident.* EPJ Nuclear Sciences & Technologies, 7 (2021): 1.

12 Schneier B. *Secrets and Lies: Digital Security in a Networked World*; Gutmann P. *Unsolvable Problems in Computer Security.*

ГЛАВА 11. Алхимия, а не наука

1 Coursey B. M. *The National Bureau of Standards and the Radium Dial Painters*. Journal of Research of the National Institute of Standards and Technology, 126 (2022).

2 *Elon Musk FULL INTERVIEW with Tucker Carlson (MUST WATCH).* April 23, 2023, 13:25, *youtube.com*.

3 Pace B. *Debate on Instrumental Convergence between LeCun, Russell, Bengio, Zador, and More*. LessWrong, October 3, 2019. Еще Лекун один раз обсуждал проблему инструментальной сходимости в комментариях к публичной ветке в Фейсбуке в 2019 году.

4 LeCun Y. *Calm down. Human-level AI isn't here yet*. X, March 19, 2023.

5 LeCun Y. *Because they would have no desire to do anything else.* X, March 20, 2023.

6 LeCun Y. *No. My benevolent defensive AI will be better…* X, March 20, 2023.

7 LeCun Y. *We can design AI systems to be both superintelligent and submissive to humans.* X, May 4, 2023.

8 McCarthy J. et al. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. August 31, 1955, *jmc.stanford.edu*.

9 Dowling S. *What Are the Odds of a Successful Space Launch?* BBC, May 31, 2023; Anand Rao T. H. *A Season Marred by Setbacks*

in Space Missions. Centre for Air Power Studies, July 25, 2024, *capsindia.org*.
Одно эмпирическое правило в ракетостроении гласит, что при первом или втором запуске ракеты нового типа шансы на взрыв составляют примерно 30 %. Регулярно взрываются даже ракеты с отработанными конструкциями. Средний показатель в истории превышает 8 % (впрочем, в последние пару десятилетий он ближе к 6 %). Даже у *пилотируемых* полетов показатель отказов за всю историю выше 1 % (в последние пару десятилетий — около 0,79 %).

10 Doman M., Sveen B. *AI's Dark In-joke*. ABC News, July 14, 2023; METR (Model Evaluation & Threat Research). *Q&A with Geoffrey Hinton*. June 27, 2024, 38:07, *youtube.com*.
Хинтона, похоже, дезинформировали насчет степени уверенности Юдковского в катастрофе, поскольку он назвал величину 99,999 %. Хотя мы, авторы, и считаем катастрофу наиболее вероятным исходом по умолчанию, пять девяток — это абсурдный уровень уверенности, под которым ни один из нас не подпишется.

11 Doman M., Sveen B. *AI's Dark In-joke*.

12 METR. *Q&A with Geoffrey Hinton*.

13 *Nearly Half of OpenAI's AGI Safety Researchers Resign Amid Growing Focus on Commercial Product Development.* Benzinga, August 28, 2024; Goldman S. *Exodus at OpenAI: Nearly half of AGI Safety Staffers Have Left, Says Former Researcher.* Fortune, August 28, 2024; Hashim S. *OpenAI Employee Says He Was Fired for Raising Security Concerns to Board.* Transformer (blog), June 4, 2024; Metz R., Ghaffary S. *OpenAI Dissolves High-Profile Safety Team after Chief Scientist Sutskever's Exit.* Bloomberg, May 17, 2024; Samuel S. *'I Lost Trust': Why the OpenAI Team in Charge of Safeguarding Humanity Imploded*. Vox, May 18, 2024.
Леопольд Ашенбреннер и Павел Измайлов были уволены, предположительно, за утечку данных компании. По утверждению Ашенбреннера, его уволили за то, что он поднял вопросы безопасности перед советом директоров. Руководители группы Ян Лейке и Илья Суцкевер ушли сами — наряду с Уильямом Сондерсом, Райаном Лоу, Яном Хендриком Кирхнером, Коллином Бернсом, Джеффри Ву, Джонатаном Уэсато, Стивеном Биллсом, Юрием Бурдой, Тодором Марковым и соучредителем Джоном Шульманом. По состоянию на начало 2025 года Лео Гао и Боуэн Бейкер все еще работали

в *OpenAI*, но группу, занимавшуюся сверхсогласованием, расформировали.

14 Haselton T., Goswami R. *OpenAI Co-founder Ilya Sutskever Announces His New AI Startup, Safe Superintelligence*.

15 Robison K. *OpenAI Researcher Who Resigned over Safety Concerns Joins Anthropic.* The Verge, May 28, 2024.

16 Jābir ibn Ḥayyān. *Kitāb al-Aḥjār 'alá Ra'y Balīnās*. 8th–9th century, trans. Haq S. N. in *A Critical Study of Jābir ibn Ḥayyān's Kitāb al-Aḥjār 'alá Ra'y Balīnās.* Thesis, University College London, 1990, *discovery.ucl.ac.uk.*
Джабир ибн Хайян, известный как отец арабской химии, — алхимик (или группа алхимиков), который добился успехов в очистке металлов. По поводу превращения свинца в золото он писал:

> Что касается преобразования тел из одного состояния в другое, более высокое или более низкое, то это, согласно нашей доктрине, внешнее и внутреннее, ибо в действительности это то, чем являются внешнее и внутреннее. Причина в том, что все составляющие всех вещей следуют круговому образу перемен.
> У любого тела внешнее явно, тогда как внутреннее скрыто, и именно в последнем кроется польза. Например, внешнее для свинца — это дурно пахнущий свинец, и это заметно для всех людей. Но его внутреннее — золото, которое скрыто. Однако если его извлечь, то и внутреннее, и внешнее для свинца станут явными.

ГЛАВА 12. Не хочу показаться паникером

1 Loeb A. P. *Birth of the Kettering Doctrine: Fordism, Sloanism and the Discovery of Tetraethyl Lead.* Business and Economic History, 24, no. 1 (1995): 72–87.

2 Kitman J. L. *The Secret History of Lead.* The Nation, March 2, 2000; HOT ROD Staff. *Living with Unleaded: Here's How Your Classic Musclecar or High-Perf Street Machine Can Safely Kick the Habit*. Hot Rod, March 1987.
Самой многообещающей альтернативной добавкой был, видимо, этанол, однако для двигателей, использовавших альтернативы свинцу, требовались более прочные клапаны и седла клапанов.

3 McFarland M. J. et al. *Half of US Population Exposed to Adverse Lead Levels in Early Childhood.* Proceedings of the National Academy of Sciences, 119, no. 11 (2022).

4 Higney A. et al. *The Lead-Crime Hypothesis: A Meta-analysis*. Regional Science and Urban Economics, 97 (2022).

5 Loeb A. P. *Paradigms Lost: A Case Study Analysis of Models of Corporate Responsibility for the Environment.* Business and Economic History, 28, no. 2 (1999); Kovarik W. *ETHYL: The 1920s Conflict over Leaded Gasoline and Alternative Fuels.* Presented at the American Society for Environmental History Annual Conference, Providence, RI, March 26–30, 2003.

6 Kovarik B. *Charles F. Kettering and the 1921 Discovery of Tetraethyl Lead.* International Fuels & Lubricants Meeting & Exposition, October 1, 1994 (1999), *environmentalhistory.org*.

7 Edelmann F. T. *The Life and Legacy of Thomas Midgley Jr.* Papers and Proceedings of the Royal Society of Tasmania, 150, no. 1 (2016): 45–49.

8 Там же.

9 Ord T. *The Precipice.* Grand Central Publishing, 2021. [Орд Т. *На краю пропасти*. М.: Corpus, 2023.]

> В моих оценках, приведенных выше, действительно учитывается возможность, что мы возьмемся за ум и начнем со всей серьезностью подходить к этим рискам. Будущие риски часто оцениваются исходя из предпосылки, что жизнь продолжится "в привычном режиме": что уровень нашего беспокойства и объем ресурсов, выделяемых на работу с рисками, останутся такими же, как сегодня. Если бы я допустил, что жизнь пойдет в привычном режиме, я бы присвоил рискам значительно более высокую вероятность.

10 Thornhill J. *How Fatalistic Should We Be on AI?* Financial Times, February 22, 2024.

11 METR. *Q&A with Geoffrey Hinton*.

12 Sunak R. *Prime Minister's Speech on AI: 26 October 2023.* United Kingdom of Great Britain and Northern Ireland, October 26, 2023, *gov.uk*.

13 Plokhy S. *Chernobyl: The History of a Nuclear Catastrophe*.

14 Titanic Inquiry Project. *British Wreck Commissioner's Inquiry | Day 6 | Testimony of Charles Joughin (Chief Baker, SS Titanic).* May 10, 1912, *titanicinquiry.org*.
Главный пекарь "Титаника" Чарльз Джокин описал, как сложно

было найти женщин и детей, желающих сесть в спасательные шлюпки, и как он и другие силой заставляли некоторых пассажиров спускаться в шлюпку.

15 *Elizabeth Weed Shutes: Titanic Survivor*. Encyclopedia Titanica, February 1, 2018.

16 Lord W. *A Night to Remember*. Penguin Books, 1976.

17 Tangalakis-Lippert K. *Elon Musk Says There Could Be a 20% Chance AI Destroys Humanity — but We Should Do It Anyway.* Business Insider, March 31, 2024; The Logan Bartlett Show. *Anthropic CEO on Leaving OpenAI and Predictions for Future of AI.* October 6, 2023, 1:38:35, *youtube.com.*

18 Plokhy S. *Chernobyl: The History of a Nuclear Catastrophe*.

19 Hinton G. *In the NYT today, Cade Metz implies that I left Google so that I could criticize Google.* X, May 1, 2023.

20 Garling C. *Andrew Ng: Why 'Deep Learning' Is a Mandate for Humans, Not Just Machines.* Wired, May 5, 2015.

21 Cotra A. *Draft Report on AI Timelines.* Alignment Forum, September 18, 2020.

22 Altman S. *The Intelligence Age*. September 23, 2024, *ia.samaltman.com*; Hern A. *Elon Musk Predicts Superhuman AI Will Be Smarter Than People Next Year.* The Guardian, April 9, 2024; Amodei D. *Machines of Loving Grace*.

23 LeCun Y. *I said that reaching Human-Level AI 'will take several years if not a decade'*. X, October 16, 2024.

24 CNBC Television. *Anthropic CEO: More confident than ever that we're 'very close' to powerful AI capabilities.* January 21, 2025, 2:05, *youtube.com*; Hern A. *Elon Musk Predicts Superhuman AI Will Be Smarter Than People Next Year;* Anderson L. *Elon Musk: A Machine Tasked with Getting Rid of Spam Could End Humanity.* Vanity Fair, October 8, 2014.

Генеральный директор *xAI* спрогнозировал появление ИИ "умнее любого человека" к концу 2025 года. Генеральный директор *Anthropic* предсказал создание ИИ, превосходящего "почти всех людей почти во всех задачах", примерно в 2027 или 2028 году. Оба они по отдельности признали, что возможности автоматизированных исследований могут вызвать взрывное развитие ИИ.

25 Werner J. *AI Superpowers & Global Treaties*. Forbes, February 19, 2025.

ГЛАВА 13. Остановите всё

1 Scher A. *Mechanisms to Verify International Agreements about AI Development*. MIRI Technical Governance Team, December 2, 2024, *techgov.intelligence.org.*

2 United Nations Office for Disarmament Affairs. *Treaty on the Non-Proliferation of Nuclear Weapons (NPT)*.

3 Burns K. *War Production.* The War | PBS, May 21, 2021; *WWII: Mobilization by Country 1937–1945.* Statista, August 9, 2024; *WWII: Annual Tank and Self-propelled Gun Production 1939–1945, by Country.* Statista, August 6, 2024; *WWII: Annual Production of Major Naval Vessels 1939–1945, by Country.* Statista, August 6, 2024.

4 Chmielewski K. *Casualties of World War II | Statistics, by Country, & Total.* Encyclopaedia Britannica, August 31, 2023.

5 Cumming B. *US House of Representatives Call for Legal Liability on Deepfakes*. Future of Life Institute, October 16, 2024; *Ban Deepfakes*. bandeepfakes.org, n.d.

6 *SB-1047 Safe and Secure Innovation for Frontier Artificial Intelligence Models Act*. leginfo.legislature.ca.gov

7 Существуют и исключения — организации, призывающие к мораторию, такие как *ControlAI*, *PauseAI* и *StopAI*.

ГЛАВА 14. Пока есть жизнь, есть и надежда

1 ControlAI. *Campaign Statement.* February 6, 2025, *controlai.com/statement*.
Сунак — не единственный британский политик, который обращает на это внимание. По состоянию на начало 2025 года десятки парламентариев Великобритании подписали заявление, где говорится: “Системы искусственного сверхинтеллекта поставят под угрозу национальную и глобальную безопасность”.

2 *Global AI Governance Initiative — The Third Belt and Road Forum for International Cooperation*. People’s Daily Online, n.d., *beltandroadforum.org*.

3 *Japan PM Vows to Lead Setting Up Int’l AI Rules through New Framework*. Kyodo News+, May 3, 2024, *english.kyodonews.net*; Alper A. *UN Adopts First Global Artificial Intelligence Resolution*. Reuters, March 21, 2024, *reuters.com*; Bennett J.T. *Biden Warns about AI’s ‘Risks,’ Forces of ‘Retreat’ in Final UN Address.* Roll

Call, September 24, 2024, *rollcall.com*; *Chinese Vice Premier Ding Xuexiang at World Economic Forum.* C-SPAN, January 21, 2025, 38:01, *c-span.org*.
В 2023 году страны "Большой семерки" запустили "Хиросимский процесс по искусственному интеллекту" — первую в мире международную структуру, касающуюся ИИ. В мае 2024 года премьер-министр Японии Фумио Кисида сказал:

> Давайте сотрудничать как нации, объединенные общей целью, чтобы реализовать возможности и устранить риски, вызванные искусственным интеллектом, и работать над достижением безопасного, защищенного и надежного ИИ.

Первая (не имеющая обязательной силы) резолюция Организации Объединенных Наций была принята в марте 2024 года. Посол США при ООН Линда Томас-Гринфилд похвалила ее:

> Сегодня все 193 члена Генеральной Ассамблеи ООН высказались заодно и вместе решили управлять искусственным интеллектом, а не позволять ему управлять нами.

В сентябре 2024 года президент США Джо Байден высказался о международном сотрудничестве по ИИ в своем заключительном обращении к Организации Объединенных Наций:

> Прежде всего, как нам, международному сообществу, управлять искусственным интеллектом, пока страны и компании мчатся к неизвестным рубежам? Нам необходимы столь же неотложные усилия для обеспечения безопасности, защиты и надежности в сфере ИИ.

На Всемирном экономическом форуме в январе 2025 года вице-премьер КНР Дин Сюэсян сказал:

> Если мы позволим продолжать безрассудную конкуренцию между странами, то увидим "серого носорога". И что с этим делать? Я думаю, нам нужно обратиться к истории. Например, к нашему опыту управления рисками: ядерными рисками, биологическими и угрозами безопасности. <…> Мы готовы в рамках Организации Объединенных Наций и ее структур активно участвовать в диалоге со всеми соответствующими международными организациями и со всеми странами, чтобы обсудить формулировки жестких правил. И гарантировать, что ИИ-технологии станут пещерой сокровищ Али-Бабы, а не ящиком Пандоры.

"Серый носорог" — это тип риска, противопоставляемый "черному лебедю". "Черный лебедь" — это событие с низкой вероятностью и серьезными последствиями. "Серый носорог" — это событие с высокой вероятностью и серьезными последствиями, которое люди тем не менее склонны игнорировать или недооценивать.

4 AI Policy Institute. *Overwhelming Majority of Voters Believe Tech Companies Should Be Liable for Harm Caused by AI Models, Favor Reducing AI Proliferation and Law Requiring Political Ad Disclose Use of AI.* September 23, 2023.
Смотрите основные результаты и таблицы в разделе *About the Poll* ("Об опросе").

5 Perrigo B. *Exclusive: The British Public Wants Stricter AI Rules Than Its Government Does*. TIME, February 6, 2025.

CORPUS 944

Научно-популярное издание

ЭЛИЕЗЕР ЮДКОВСКИЙ, НЕЙТ СОАРЕС

ЕСЛИ КТО-ТО ЕГО СОЗДАСТ — ВСЕ ПОГИБНУТ

ПОЧЕМУ СВЕРХЧЕЛОВЕЧЕСКИЙ **ИИ** УНИЧТОЖИТ НАС ВСЕХ

Главный редактор и ответственный за выпуск Варвара Горностаева

Художник Андрей Бондаренко

Редактор Алёна Якименко

Научный редактор Сергей Николенко

Технический редактор Наталья Чернышева

Корректор Александра Тромбова

Верстка Марат Зинуллин

Общероссийский классификатор продукции
ОК-034-2014 (КПЕС 2008);
58.11.1 — книги, брошюры печатные

Подписано в печать 11.05.2026. Формат 60 × 90 1/16
Бумага книжная кремовая пухлая. Гарнитура *OriginalGaramondC*
Печать офсетная. Усл. печ. л. 16,00
Тираж 4000 экз. Заказ №

Отпечатано в соответствии с предоставленными материалами
в ООО "ИПК Парето-Принт", 170546, Тверская область,
Промышленная зона Боровлево-1, комплекс № 3А
www.pareto-print.ru

Охраняется законом РФ об авторском праве. Воспроизведение всей книги или любой ее части воспрещается без письменного разрешения издателя. Любые попытки нарушения закона будут преследоваться в судебном порядке.

Произведено в Российской Федерации в 2026 г.
Изготовитель — ООО “Издательство АСТ”

ООО “Издательство АСТ”
129085 г. Москва, Звездный бульвар, д. 21, строение 1, комната 705,
пом. I, 7 этаж
Контактный адрес электронной почты: ask@ast.ru
Интернет-магазины:
https://www.chitai-gorod.ru/ и https://www.bookvoed.ru/

“Баспа Аста” деген ООО
129085, Мәскеу қ., Звёздный бульвары, 21-үй, 1-құрылыс,
705-бөлме, I жай, 7-қабат
Біздің электрондық мекенжайымыз: ask@ast.ru
Интернет-дукендер:
https://www.chitai-gorod.ru/ және https://www.bookvoed.ru/

Импортер в Республику Казахстан ТОО “РДЦ-Алматы”
Дистрибьютор и представитель по приему претензий на продукцию
в Республике Казахстан: ТОО “РДЦ-Алматы”

Қазақстан Республикасындағы импорттаушы “РДЦ-Алматы” ЖШС
Қазақстан Республикасында дистрибьютор және өнім бойынша
арыз-талаптарды қабылдаушының өкілі “РДЦ-Алматы” ЖШС

050039 Алматы қ., Домбровский көш., 3 “а”, литер Б, офис 1
Тел.: +7 (727) 251-5989, 90, 91, 92, факс: +7 (727) 251-5812, доб. 107
E-mail: RDC-Almaty@eksmo.kz
http://www.book24.kz

Өнімнің жарамдылық мерзімі шектелмеген

По вопросам оптовой покупки книг обращаться по адресу:
123112 г. Москва, Пресненская наб., д. 6, строение 2,
БЦ “Империя”, тел.: +7 (499) 951-6000, доб. 574
E-mail: opt@ast.ru